这才是心理学

看穿伪心理学的本质

【加】基思·斯坦诺维奇（Keith E. Stanovich）著　窦东徽 刘肖岑　译

How to Think Straight About Psychology



10th Edition 第10版

一本为心理学去伪存真、有态度的、让你真正了解什么才是心理学的经典著作

畅销三十余年、长踞亚马逊心理类图书前100名

中国人民大学出版社

VISIT...

LANZAROTE Caliente.com

版权信息

书名：这才是心理学：看穿伪心理学的本质

作者：（加）基思·斯坦诺维奇

出版社：中国人民大学出版社

出版日期：2015-8-1

ISBN：978-7-300-21872-4

价格：49.00元

目录
CONTENTS

序言
推荐序 心理学有什么不同之处
译者序
序言
Chapter 1 心理学充满生机：在科学阵营里左右逢源
弗洛伊德问题
现代心理学的多样性
科学的统一性
那么，什么是科学
心理学和世俗智慧："常识"的误区
心理学是一门年轻的科学
小结
Chapter 2 可证伪性：如何挫败头脑中的小精灵
理论和可证伪性标准
科学中的错误：逼近真理
小结
Chapter 3 操作主义和本质主义："但是，博士，这到底是什么意思？"
为什么科学家不是本质主义者
心理学领域的操作性定义
小结
Chapter 4 见证和个案研究证据：安慰剂效应和了不起的兰迪
个案研究的地位
为什么见证叙述毫无价值：安慰剂效应
"鲜活性"问题
见证为伪科学打开方便之门
小结
Chapter 5 相关和因果：用"烤箱法"避孕
第三变量问题：戈德伯格与糙皮病
方向性问题
选择性偏差
小结
Chapter 6 让一切置于控制之下：聪明汉斯的故事

斯诺与霍乱
比较、控制和操纵
小结
Chapter 7 不像是真实生活的心理学实验与“人为性”批评
为什么自然性并非总是必要的
心理学理论的应用
小结
Chapter 8 避免爱因斯坦综合征：聚合性证据的重要性
关联性原则
科学共识
不要对矛盾数据感到绝望
小结
Chapter 9 打破“神奇子弹”的神话：多重原因的问题
交互作用
单一原因解释的诱惑
小结
Chapter 10 人类认知的阿喀琉斯之踵：概率推理
“某某人”统计学
概率推理以及对心理学的误解
有关概率推理的心理学研究
小结
Chapter 11 偶然性在心理学中扮演的角色
试图解释偶然性事件的倾向
偶然性和心理学
接受错误以减少错误：临床预测与统计预测
小结
Chapter 12 不招人待见的心理学
心理学的形象问题
心理学和其他学科
我们是自己最坏的敌人
每个人不都是心理学家吗
抵制科学心理学的根本原因
结束语

推荐序 心理学有什么不同之处

彭平凯 清华大学心理学系教授、系主任

心理学是一门很容易让人误解的学科。与其他学科不同，它研究的是与人民生活紧密相关的问题。有些还正好是大家都熟悉，而且经常关心的问题。比如：什么样的人容易讨人喜欢？什么样的事情我们记忆深刻？什么样的事情让人高兴（或痛苦）？为什么男人比女人更爱聊政治时事？人为什么要自杀？意念能不能被植入梦中？等等……正因为大家关心这些问题，人们就会有自己的分析，自己的证据，得出自己相信的结论。很多时候，这些自觉的结论与心理学家的研究结论并不完全一致，比如说，我们心理学家就发现，青梅竹马的婚姻很难成立，婴儿并不是有奶便认娘，性格并不决定一个人的命运，等等……每当矛盾、怀疑、迷惑甚至气愤产生的时候，我们到底是该相信自己的直觉、经验和常识呢，还是该相信心理学的研究、证据和知识呢？

不幸的是，大多数的心理学教科书只满足告诉大家心理学的研究、发现和知识，但从不说明为什么这些研究、发现和知识是值得我们关注和信任的。这些书的作者好像都在假设每一位读者思考起问题来都像心理学家一样，相信和理解心理学的研究、发现和知识。幸运的是，基思·斯坦诺维奇教授写了这本这本《这才是心理学》，一本“与众不同的心理学”教科书。他总结了心理学家的职业特质，让每一位读者都有机会去理解我们心理学家是如何去思考、分析和解读人类的行为和心理的。他的每一章都将一个常识的、朴素的、直觉的有关人类心理的分析和思考与一个科学的、严谨的、心理学的分析和思考相对比，以帮助读者理解心理学家的分析逻辑和研究思路。

我个人觉得，在斯坦诺维奇阐述的心理学特质中，有两点应该是区分我们心理学家和其他人（包括其他领域的科学家）与众不同的地

方。

其一是我们心理学的批判性思维习惯，也就是说，心理学不相信个人的智慧，更相信科学的方法，而科学方法的本质是证伪，即对我们的经验、常识和直觉，产生怀疑、挑战和批评。从原则上而言，心理学家不怕犯错误，但害怕以假乱真。心理学家也不相信那些能回答所有问题的绝对真理，但相信对所有问题应该有一个相对正确的答案。心理学家从不相信个案和例子，因为其随机性和主观性太过明显，但我们愿意相信大样本基础上的科学研究发现。我们希望听到动听的心理故事，但更愿意看到众多心理学观察的数据和总结。

其二是我们心理学的概率性思维习惯。我们和很多自然科学家思考方式不同之处就在于我们更容易相信，任何人类的行为都是概率性的表现，也就是说，它有一定的不确定性，会受到其他随机事件的影响。其实人类很多学科都是建立在概率基础之上的，比如说核物理、天体物理、生物进化、病理学、所有的社会科学（经济学、社会学、政治学等）等等，它们都不可能准确预测每一个研究对象的具体活动，但都对整体的事物活动规律有很好的描述和预测。只不过我们心理学家更愿意承认而且强调我们学科的不确定性而已。

总之，我很高兴地看到这本书在中国再版。杨中芳老师的早期译本是我很喜欢向心理学爱好者推荐的一本心理学教科书。而新版的译本，尤其令我兴奋，因为它是由我欣赏的两位年轻同事——东徽和肖岑，下工夫，花时间，认真翻译完成的。两位既是同行，也是夫妻，应该是我们心理学界不鲜见的学术伴侣之一。这可能也是我们心理学界与众不同的地方吧。

译者序

此刻诸位手中拿的正是斯坦诺维奇教授的《这才是心理学：看穿伪心理学的本质》（How to think straight about psychology）第10版中译本。记得我的博士后合作导师、清华大学心理学系主任、加州大学伯克利分校终身教授彭凯平老师曾经说过：在美国，如果一本心理学教科书能够再版3次以上，这本书就堪称经典教材了。按照这一标准，斯坦诺维奇教授这本书堪称经典教材中的翘楚，无怪乎能够成为诸多国家高校心理学本科专业学习的指定书目。

斯坦诺维奇教授曾感慨，写作最新一版的初衷，与写第一版时无异。此感慨中有不忘初心方得始终的赤子情怀，但也有些许的无奈：心理学被大众误解的状况多年来并未有实质性的转变，心理学这些年的蓬勃发展也没有使这一情况有根本性的好转。对于这一点，相信每一位学习和从事心理学的人都有所体会。

时至今日，每个学心理学专业或从事心理学工作的人仍然会被人问到三个问题，第一个通常是：“你知道我现在心里在想什么吗？”名列榜眼的问题则是：“你会催眠和解梦吗？”排名第三的问题则一般是：“什么，你们心理学还做实验？”这类对于问话人来说再正常不过的问题，却总能让这些学心理学的人哭笑不得。

一个真正称职的心理学工作者不应回避或敷衍这类问题，正确的做法是直面它们并认真分析其中的意义，给予人们正确的解答。从最常被问到的这三个问题中，我们便能以管窥豹地看出公众对于心理学的一些常见的误解和迷思。

第一个问题所揭示的事实是，心理学在大众心目中被万能化了。所谓万能化，是指心理学研究对象和功能的边界被无限地夸大了。首先，许多人认为心理学无所不包，星座、属相、血型、玄学、人生哲

理及各种心灵鸡汤都被认为是心理学的范畴——实际上，心理学研究虽然涉及广泛的人的客观行为和心理现象，但作为心理学的研究对象必须满足“可检验”这一标准，超验的、超感官的问题不在科学心理学的研究之列；另外，并不是说心理学不能研究血型、星座和超感知，实际情况是，这类现象和理论在历史上都曾经是心理学的研究对象，但已被各种科学的方法所证伪并从心理学的研究对象中剔除了出去。还有一种观点认为，心理学是无所不能的，学了心理学就能读心、能算命，甚至能千里之外控制他人的大脑并操纵其行为，还兼具其他种种神奇的功效。我们说心理学很神奇，是因为它能在一定程度上解释心理现象和预测行为，并能提供一些行之有效的干预方法，但是，科学心理学所有的预测和解释都基于客观数据，所有的结论都具有概率的性质，有一定的适用范围和条件，干预方案也必须符合客观的心理和行为规律。因此没有能够完美解释一切现象的心理学理论，也没有超出人类经验以外的干预和应用。

第二个关于催眠的问题则揭示了另外一种对于心理学常见的误解：心理学就是弗洛伊德的精神分析，就是心理咨询。弗洛伊德这位伟大的心理学家确实让后世的心理学研究者“既爱又恨”。弗洛伊德的功绩在于，他以其非凡的工作投入和严谨的思辨构筑了一套宏大、晦涩而让人着迷的理论体系，开启了一片探索人类心灵世界的全新领域，对社会文化、艺术创作产生了不可估量的影响，并真正让世人开始了解和重视心理学；迄今为止他的理论仍然广泛应用于心理治疗和干预的诸多领域。今天人们都能熟练用“焦虑”、“压抑”、“潜意识”这些词来描述或解释分析自己的心理状态，也是拜弗洛伊德和他的精神分析学派所赐。说到心理学家，或许很多人不知道冯特和斯金纳，但没有人会不知道弗洛伊德。然而，弗洛伊德的盛名也带来了一系列问题，最主要的是两点：一是他的光芒掩盖了其他领域心理学研究者的工作和成就，二是弗洛伊德所构建的这些精巧的理论很难用实证的方法加以验证，这也为后来众多伪心理学和虚假治疗手段的滋生和发展留下了一道后门。

最后一个问题也是最关键的——公众不了解主流的科学心理学已经在实证主义的道路上行进了很远。心理学是一门科学，必然要遵循科学的标准，即研究的必须是实证可解的问题；方法上要遵循系统的实证主义；研究结论是公开知识，即能够被重复验证（可再现性），并能经过同行评议。这三条标准也是区分科学心理学和其他伪心理学的分水岭。心理学的研究因此也与操作定义、实验组和控制组、变量控制、数据统计等词汇联系起来，而不是个人化的体验及感悟、头脑风暴或纯粹的逻辑推导。

那么，是什么阻碍了公众了解和认知真正的科学心理学？原因应该是多方面的。首先，对心理学的渴求使得大众对心理学产生了特殊的期待。心理现象和心理问题与每个个体戚相关，人们试图通过心理学解释这些问题和现象，并从中获得行之有效的帮助和建议。在这方面，针对个体的精神分析和治疗技术有一定的优势，但其他大多数心理学研究则针对的是群体的普遍行为规律，偏重于解释和预测，其研究结论都具有概率性和领域特异性，因此无法有针对性地、面面俱到地解决个体所有的心理问题。这有时难免会让一部分抱着“求医问药”的心态来了解心理学的人感到失望。此外，以“求医问药”为动机接触心理学的人，也往往对心理学的其他领域缺乏关注。

第二，术语体系产生的阻隔。心理学作为一门科学，为了让研究者之间形成共识并利于重复验证，产生了一套严谨的术语体系，这一点和其他科学学科并无二致。但不同的是，没有人会因为不懂量子力学的各种晦涩的术语而鄙视物理学，但他们却会因不能忍受心理学术语所造成的隔膜而对科学心理学产生拒斥，继而转向了那些通俗易懂但科学性无法得到保证的理论或方法。究其原因，还是因为心理问题和现象与每个人的关系太为紧密了，最近的需要总是要以最快捷的方式介入和解决，这就构成了快速实用和科学研究之间的一对矛盾。

第三，心理学许多基础研究确实和大众生活有一定的距离。在所有的科学门类中，基础研究转化为实际应用需要时间，有的研究发现甚至要几十年之后才能体现出其应用价值，心理学的基础研究向应用的转化也概莫能外。这种滞后性有时会让人产生一种科学心理学研究毫无意义的错觉。

最后一个，也是非常重要的一个原因，就是大众媒体在科学心理学的传播方面做得并不够好。随着心理学的热度持续升温，许多电视、电台、网站和出版商和纸媒也不断推出与心理学相关的节目、书籍或专栏，但由于种种原因（如专业限制、商业考虑等等），最终呈现在大众面前并得到广泛传播的往往并不是科学的心理学，而是包装精美的伪心理学；有一些所谓的“心理学家”甚至在媒体平台上用错误的理论误导大众。与之相对应的是，真正的心理学家和专业书籍无人问津，而星座、血型、养生以及各种未经实证检验的古怪疗法却打着心理学的旗号招摇过市，让心理学蒙受了许多质疑和指责。

基于以上种种，如何让公众了解真正科学的心理学变得十分必要。市场上有关心理学的书籍很多，但对象分化的情况也很突出：针对心理学专业学生和心理学工作者的专业书籍能够提供很多有用的信息和知识，但有专业门槛的隐形限制；而针对普通读者的非专业书籍相对通俗但在信息质量方面良莠不齐。斯坦诺维奇教授所写的这本

《这才是心理学：看穿伪心理学的本质》的定位则兼顾心理学的初学者和对心理学感兴趣的更广大的读者群体，从质量来说应为此类心理学入门读物中之翘楚。在这本20万字左右的书里，作者以生动而严谨的笔触告诉了读者什么才是真正的心理学。书中重申了科学心理学诸多重要标准和核心理念，澄清了有关心理学的种种误解和迷思，有破有立，言之凿凿；理论讲述与精彩的实验案例交相呼应，集科学性和趣味性于一体，十分耐读；有些犀利的论点足以对读者原有的知识信念构成挑战，但这种不安很快就会被知识重构的提升感和思辨的乐趣所取代。

这本书先前的版本曾经由杨中芳老师翻译，已使很多读者从中受益。从第8版开始到今天的第10版，由我与刘肖岑老师共同翻译，对这3版的翻译也让我们处于知识更新的喜悦中。书已近付梓，但译者水平所限，难免有不周或纰缪之处，还请广大读者给予指正。

本书第10版在第9 版的基础上有所更新，删除了一些比较陈旧的文献和被最新的研究证明是有争议的结论，同时补充了最新的研究结论和案例，反映了所涉领域最新的进展。此外，作者对一些评述文字的说法和措辞进行了调整，使之更切合所论述的主题。全新升级之后，这一版不仅原汁原味地保留了之前版本的精髓，同时在内容上更为丰富和具有时效性。

我们要感谢北京师范大学心理学院邹泓教授，华东师范大学心理系桑标教授，美国加州大学伯克利分校教授、清华大学心理学系教授彭凯平，山东师范大学心理系张文新教授，首都师范大学心理系方平教授，以及中央财经大学社会发展学院辛自强教授，在本书翻译过程中给予的指点和帮助，以及一直以来在学术和思想方面的引导和教诲。同时，第10版的顺利出版，与人民大学出版社各位编辑认真严谨的工作是分不开的，在此一并致谢。

要感谢的还有：中央财经大学的张杰、姜涛、高霖宇、丁志宏、赵然、张红川、黄四林、翁学东、冯源、辛志勇、于泳红、孙铃、苑媛、张梅、侯佳伟、马敏、汪波、赵娜；首都师范大学的王建平、李莉、田汉族、许晓晖、于开莲、高维华、严冷、刘昊、梁九清、黄翯青、夏婧、钱国英、周楠、孟繁华、丁锦红、郭春彦、王争艳、王异芳、梁熙等诸位老师，以及我们的学生陈婷、钱天月、张玉洁、杜宇、弓仲冬、谭洁、张惠、张缙、郭晓红、王娟和齐小琳。感谢你们在本书翻译、出版过程中提供的各种帮助和支持。

美国心理学会前任主席、积极心理学创始人马丁·赛利格曼教授曾用一个、两个和三个词形容心理学现在的状态，分别

是：“good”（好），“not good”（不好）和“not good enough”（还不够好）。的确，心理学是一门蓬勃发展的科学，当前虽然有不尽如人意的地方，但我们坚信，伴随着每一位科学心理学工作者的不懈努力，心理学必将变得更好，帮助人们实现更大的福祉。心存此志，无远弗届。

窦东徽 刘肖岑

序言

有这样一门尚不为大多数人所知晓的知识，它涉及人类行为和意识的不同形式，可以被用来解释、预测和控制人类的行为。学习这门知识的人能够更好地理解他人，对决定他人想法和行为的原因有更加全面和精确的认识。

你可能想不到，这门不被知晓的知识就是心理学。

当我说心理学仍不为人所知时意味着什么？你一定认为此话不能当真。如今，书店里充斥着大量标题为心理学的书籍，电视和广播脱口秀定时播放关于心理学的主题，报纸和杂志也辟有心理学专栏，怎么能说心理学无人知晓呢？但从某种关键的意义上来讲，心理学确实仍是一个不为人知的知识领域。

尽管心理学似乎得到了众多媒体的关注，但是心理学这个知识体系的绝大部分内容仍不为公众所知。经由大众媒体传播的“心理学”知识在很大程度上只是一种幻象。很多人不知道他们在书店里看到的大部分所谓心理学书籍，都是由一些在心理学界根本站不住脚的人写的；很多人也不知道，多数在电视上号称心理学家的人，根本得不到美国心理学会（American Psychological Association, APA）和美国心理协会（Association for Psychological Science, APS）的承认；他们更不知道，大多数频频亮相的心理学“专家”，其实对心理学领域的知识积累没有作出丝毫的贡献。

媒体这种对于“心理学”话题的浅薄关注，不仅向公众传递了许多错误信息，还遮蔽了心理学领域中真正的、不断发展的知识基础。公众不能肯定到底哪些是心理学，哪些又不是，也不知道该怎样独立对有关人类行为的主张作出判断。更大的问题在于，很多人始终觊觎着那些要么缺乏判断能力、要么认为无法对心理学主张作出判断的公

众。后一种观点有时被称为“无一定之规”的态度，是本书要讨论的谬误之一，这种心态对社会危害极大。公众不知道，关于行为的言论是可以验证的，很多伪科学正是利用公众的这些无知建立起百万美元的产业。人们并不知道许多伪科学（例如占星术、通灵外科手术、超速阅读、生物节律、接触治疗、潜意识自助录音带、辅助沟通和灵媒侦探等）所宣称的主张，其实早已被证明是错误的。本书提到的这些伪科学产业的存在，助长了媒体对伪科学进行炒作式报道的趋势。这种趋势对心理学的危害远比对其他学科的危害要大，理解个中缘由，是学习如何正视心理学的一个重要环节。

本书面向的不是即将成为心理学的研究者，而是面向一个更大的读者群——心理信息的消费者。本书的读者对象是心理学初学者，以及那些在大众媒体上得知一些心理学话题、又想知道如何去判断这些信息的合理性的广大读者们。

本书不是一本标准的心理学入门教材，它没有对心理学领域已取得的研究成果进行总结。事实上，单靠到大学里选修一门心理学导论的课程，对于纠正公众心中已经被媒体误导了的对心理学的看法来说，可能并不是终极的解决方案。因为众多对心理学抱有很大兴趣的非专业人士没有时间、没有钱或是没有机会进入大学进行正规的学习。更重要的原因是，作为一名大学心理学的教师，我不得不承认我和同事们在大学里教心理学时，也总是没能引导初学者对心理学这门科学产生一个正确的认识。原因在于初级的课程设置中通常没有包含对批判性分析思维技巧的训练，而这正是本书讨论的焦点所在。作为教师，我们常常只会忙于 将“研究发现”塞入教学内容当中。每次我们在和学生讨论到诸如媒体眼中的心理学等稍微偏离教学大纲的话题时，都会感到有些内疚，并开始担心自己会不会因为跑题而不能在学期结束前完成所有授课内容。

现在的心理学导论类教科书通常都有600～800页厚，并且引用了数百篇已经发表的文献。当然，包含如此丰富的材料并没什么错，至少它反映了心理学知识在不断增长。然而不幸的是，负面效果也同时存在。教师们常常只忙于给学生灌输一大堆的理论、事实和实验，而没时间去关注那些会被学生带入心理学研究的基本问题及错误观念。这主要是因为教师们（包括入门类教科书的作者）想当然地认为，只要学生接触了足够多的心理学研究，自然就能从中推导出问题的答案。简而言之，他们希望学生可以从对心理学各领域实证研究的讨论中，自行挖掘出各类问题的隐含答案，但是这类希望往往都落空了。到这门课的最后复习阶段或学期结束之前，教师们才无比震惊和沮丧地发现，学生提出的一些问题及说法，是他们在课程开始第一天就应

该提出来讨论的，而不是在14周之后。比如："既然心理学实验不同于现实生活，那么它们能告诉我们什么呢"、"心理学能像化学那样成为一门真正的科学吗"、"可是，我在电视上听一位心理大师讲的正好与我们教科书上说的相反"、"我认为这个理论相当愚蠢，因为我弟弟的行为和这个理论所说的截然相反"、"心理学不过是些常识"、"每个人都知道什么是焦虑，何苦还要去定义它呢"、"心理学不过是一堆观点而已"。对于很多学生来说，仅靠思考书中的内容是无法为这些问题找到答案的。在本书中，我将对这类问题和说法背后的误区进行澄清。

不幸的是，研究发现，普通的心理学入门课程并不能有效地纠正初学者对心理学所持的诸多误解（Keith Beins, 2008; Kowalski Tayloer, 2009; Standing Huber, 2003）。这一不幸的事实赋予了本书以合理性。心理学学生需要批判思维方面的正确指导，批判性思维能够使其对心理学信息作出独立的评估。

即使若干年后学生们不再记得心理学入门课程中的内容，但他们仍然可以运用本书所涉及的基本原理去判断心理学的主张。即使埃里克森（Erikson）的人生发展阶段论被忘得一干二净，他们也仍可以运用本书介绍的思维工具去辨别媒体中出现的心理学信息的真伪。一旦掌握，这些技能就可以成为终身受用的思维工具，帮助我们去评判各种知识主张：首先，它将使我们能够对那些看似合理的事实做出一个初步和总体的判断；其次，这些技巧提供了一些评估"专家"观点可信度的标准。因为在复杂的社会中，人们对专家观点依赖始终存在，在获取知识时，对专家观点可信度的判断就变得尤为重要了。虽然这些批判性思维技巧可运用于各个学科或知识门类，但它们在心理学领域里尤其重要，因为此领域经常被大众媒体所歪曲。

许多心理学家都对试图阻止心理学被歪曲的努力持悲观态度。虽然这种悲观并非没有道理，但是这本类似"消费者指南"式的书源自一个信念，那就是心理学家不能让这一问题成为一种自我实现式的预言。

尽管我很高兴这本书能有多次再版的机会，但令人遗憾的是，本书存在的原因仍和本人当初撰写本书第一版时完全一样。媒体对心理学的介绍一如既往地是在误导大众，而学生在开始上心理学入门课程时，还是带着与以往同样多的对于心理学的误解。正因为如此，本书后续几个版本的目标始终如前。这些目标也正在被越来越多心理学教师所共享。斯坦福大学心理学家罗杰·夏佩德（Roger Shepard）表达了与本书第一版写作初衷相同的看法："虽然大多数心理学的本科生或许不会走上学术科研的道路，但我们仍然希望他们有能力去对那些

不断出现在媒体上的片面、幼稚、混乱及夸张的所谓社会科学'发现'的报导作出判断……那些广为流传、认为可以通过片面的常识或更为糟糕的星相学之类的伪科学就能充分理解人类行为和心理现象的谬论，势必向我们持续发出挑战。"（Shepard, 1983, p.855）

批判性思维技能帮助人们更好地理解心理学的主题，以及他们周围的世界所发生的事情。本书将对此技能作一个简略的介绍。

第10版更新的内容

因为之前版本已经进行了章节重组，本书第10版在结构上没有作很大的改动，各章节的内容和顺序也保持原样。并且应评审者和读者的要求，这一版与第9版篇幅相当。我更新和重写了书中的许多例子（保留了读者反响最好的例子），用最近的研究和主题替换掉了那些过时的案例。本版最大的修订在于引用与本书所提及的各种概念和实验相关的最新研究资料。因此，大量新的引文出现在这一版中（确切来说是172条新引文），读者可以获得有关样例和概念最新的参考文献。

本书一如既往会对批判性思维技能作简略介绍，帮助学生更好地去理解心理学的主题。在过去的15年里，大学里强调批判性思维技能的呼声越来越高（Abrami et al., 2008; Sternberg, Roediger, Halpern,2006）。的确，美国一些州立大学系统已经进行了以加强批判性思维教育为目的的课程改革。与此同时，也有其他教育学学者认为，批判性思维技能不应该脱离特定的学科内容。而《这才是心理学》正好融合了这两种取向，在帮助教师教授丰富的现代心理学知识的同时传授批判性思维的技巧。

欢迎读者们将对本书的意见发送到以下邮箱：Keith.stanovich@utoronto.ca.

Chapter 1
心理学充满生机：在科学阵营里左右逢源

弗洛伊德问题

在大街上随便拦住100个人，请他们说出一个健在的或已故的心理学家的名字，然后记下他们的答案。毫无疑问，他们提到的会是菲尔博士（Dr. Phil）、韦恩·戴尔[1]（Wayne Dyer）以及其他一些“媒体心理学家”。如果我们把这类媒体和通俗心理学家排除在外，只考虑那些对心理学作出过卓著贡献的心理学家，那么这项非正式小调查的结果就几乎没什么悬念了——西格蒙德·弗洛伊德会名列榜首，B. F. 斯金纳可能会屈居次席，但得票数会远远落后于弗洛伊德。任何其他心理学家都缺乏足够的知名度来撼动这两位的地位。因此，可以说弗洛伊德和那些在媒体上频频露脸的通俗心理学共同定义了公众心目中的心理学。

弗洛伊德的声名远播，极大地影响了普通公众对心理学的理解，同时也造就了诸多认识上的误区。例如，许多刚入门的心理学学生会惊讶地发现，如果对美国心理学会（APA）会员中所有认同弗洛伊德精神分析的人数进行一下统计，他们的人数居然没有占到会员总数的10%。在另一个主要的心理学组织美国心理协会（APS）中，这一比例也绝不会超过5%（Engel, 2008）。有一本受欢迎的心理学入门教科书（Wade Tavris, 2008），在其超过700多页的篇幅中，只有15页内容提到弗洛伊德或精神分析学派，而且这15页中经常出现的是对其

的批判（“大多数弗洛伊德的观点都曾经并且现在依旧被大多数实证取向的心理学家排斥”，p.19）。

简而言之，现代心理学并没有像媒体和一些人文学科那样被西格蒙德·弗洛伊德的理论所左右，也没有被其限定。在现代心理学家所关注的大量研究主题、数据和理论中，弗洛伊德的工作只占其中极小的一部分，在这些研究和理论中占更大比重的则是5位诺贝尔奖得主所做的工作：大卫·胡贝尔（David Hubel）、丹尼尔·卡尼曼（Daniel Kahneman）、赫伯特·西蒙（Herbert Simon）、罗杰·斯佩里（Roger Sperry）和托斯腾·维瑟（Torsten Wiesel）以及美国国家科学基金的前负责人理查德·阿特金森（Richard Atkinson）的贡献。然而，这些人的名字对公众来说却相当陌生。

弗洛伊德对于现代心理学的重要性被无限地夸大了，这事儿已经足够糟糕了。更糟的是，弗洛伊德的调查方法完全不能代表现代心理学家是如何进行研究的。事实上，弗洛伊德式的研究方法彻底误导了人们对心理学研究的印象。例如，弗洛伊德并不采用控制实验，而我们将在第6章讲到，控制实验是现代心理学家“兵器库”中最有力的武器。弗洛伊德认为，个案研究足以证明理论的真实或谬误，在第4章中，我们将谈谈这一理念为何是错误的。正如一位心理治疗领域的史学家所说的那样，如果弗洛伊德本人是一名科学家，那么他所宣扬的是一门很奇怪的科学……精神分析包含理论和假设，但是缺少实证观测的方法（Engel, 2008, p.17）。

最后，弗洛伊德的工作中最大的问题是理论和研究数据的联系。正如我们将在第2章看到的，对于一个科学理论来说，理论和研究数据的联系必须满足一些标准，而弗洛伊德的理论常常不能满足这些标准（Dufresne, 2007; Engel, 2008）。简而言之，弗洛伊德根据他得到的数据（个案研究和内省）建立了一套精细的理论，而这些数据并不足以支撑此理论；他专注于构建复杂的理论构架，但并没有像许多现代心理学家那样，保证这些理论建在可靠、可重复的因果关系之上。总之，人们对于弗洛伊德式的工作太过熟悉，这严重阻碍了他们对现代心理学的正确理解。

在这一章中，我们将采用两种方法来解决“弗洛伊德问题”：首先，当我们展示现代心理学的多样性时，就能非常清楚地了解到弗洛伊德的工作所占的比重其实是很小的；其次，我们将讨论一下，在广泛而多样的心理学研究中，哪种特征是最为普遍的。有关弗洛伊德工作的那种过时的认识遮蔽了普通大众的双眼，使之无法看到现代心理学所共有的唯一而普遍的特性：用科学的方法去寻求对行为的理解。

注释

[1] 这两位是活跃于美国媒体的所谓心理学家的代表，他们这类人通常有着丰富的媒体经验，有广泛的受众，但他们没有受过严格的学术训练，不被真正的心理学学术团体认可，所传播的往往是一些似是而非的心理学知识。——译者注

现代心理学的多样性

事实上，现代心理学包含了大量不同的内容和观点。这种多样性使得作为一门学科的心理学显得不那么浑然一体。美国心理学基金会杰出教学奖得主亨利·格雷特曼（Henry Gleitman, 1981）将心理学描述为“一个松散地联合在一起的学术王国，它横跨了生物科学和社会科学两个领域”（p. 774）。心理学圈以外的人批评这种多样性。例如，人类学家克利福德·格尔茨（Clifford Geertz, 2000）抱怨道：“至少从外部来看，心理学不像是一个单一的领域，能够像通常那样被分为学派和专业。它看起来像是不同的、不相关的调查方法的大杂烩，这些调查之所以能够归为一类，是因为它们都以这样或那样的方式探究所谓心智功能。”（p.187）

心理学有着令人惊叹的广泛性和多样的调查方法，知道这一点对于理解心理学的本质至关重要。可以简单列举一些具体指标来证明这一点：美国心理学会（APA）有56个分支机构，每个分支都代表了一个特定的研究或应用领域（见表1—1）。从表1—1中，你可以看到心理学研究主题、研究背景和研究方法的丰富性和多样性。另一个大型心理学组织——美国心理协会（APS）也同样分支众多。其实，表1—1对于心理学领域的多样性的描述还是较为保守的，因为它给我们造成了一种印象，即每个分支都是一个特定的专业领域。事实上，56分支机构中的每一个都是非常宽泛的研究领域，包含更小的不计其数的分支。简而言之，要穷尽心理学领域主题的多样性是非常困难的。

表1-1 美国心理学会（APA）的分支机构

1.普通心理学（General Psychology）

2.心理学教学（Teaching of Psychology）

3.实验心理学（Experimental Psychology）

5.评价、测量和统计（Evaluation, Measurement, and Statistics）

6. 神经行为科学和比较心理学（Behavioral Neuroscience and Comparative Psychology）

7.发展心理学（Developmental Psychology）

8.人格和社会心理学（Personality and Social Psychology）

9.社会问题的心理学研究（Psychological Study of Social Issues）

10.审美、创造力及艺术心理学（Psychology of Aesthetics, Creativity, and the Arts）

11.临床心理学（Clinical Psychology）

12.应用咨询心理学（Consulting Psychology）

13.工业和组织心理学（Industrial and Organizational Psychology）

14.教育心理学（Educational Psychology）

15.学校心理学（School Psychology）

16.理论咨询心理学（Counseling Psychology）

17.公共服务中的心理学家（Psychologists in Public Service）

18.军事心理学（Military Psychology）

19.成人发展与老龄化（Adult Development and Aging）

20.应用实验和工程心理学（Applied Experimental and Engineering Psychology）

21.康复心理学（Rehabilitation Psychology）

22.消费者心理学（Consumer Psychology）

23.理论和哲学心理学（Theoretical and Philosophical Psychology）

24.行为分析（ Behavior Analysis）

25.心理学史（History of Psychology）

26.社区心理学（Community Psychology）

27.精神药理学和药物依赖（Psychopharmacology and Substance Abuse）

28.心理治疗（Psychotherapy）

29.心理催眠（Psychological Hypnosis）

30.国家心理学常务联合会（State Psychological Association Affairs）

31.人本心理学（Humanistic Psychology）

32.智力缺陷和发展性障碍（Mental Retardation and Developmenta Disabilities）

33.人口与环境心理学（Population and Environmental Psychology）

34.女性心理学（Psychology of Women）

35.宗教心理学（Psychology of Religion）

36.儿童、青少年和家庭服务（Child, Youth, and Family Services）

37.健康心理学（Health Psychology）

38.心理分析（Psychoanalysis）

39.临床神经心理（Clinical Neuropsychology）

40.心理学和法律（Psychology and Law）

41.独立从业的心理学者（Psycholigists in Independent Practice）

42.家庭心理学（Family Psychology）

43. 男女同性恋及双性恋的心理学研究（Psychological Study of

Lesbian, Gay, and Bisexual Issues）

44.少数民族的心理学研究（Psychological Study of Ethnic Minority Issues）

45.媒体心理学（Media Psychology）

46.锻炼和运动心理学（Exercise and Sport Psychology）

47.和平心理学（Peace Psychology）

48.团体心理学和团体治疗（Group Psychology and Group Psychotherapy）

49.成瘾（Addictions）

50.男性和男性化的心理学研究（Psychological Study of Men and Masculinity）

51.国际心理学（International Psychology）

52. 临床儿童心理学和青少年心理学（Clinical Child Psychology and Adolescent Psychology）

53.幼儿心理学（Pediatric Psychology）

54.药物疗法（Pharmacotherapy）

55.创伤心理学（Trauma Psychology）

注：没有分支4和11。

多样性的含义

许多人学习心理学是希望能够学到一套宏大的心理学理论，以此来概括和解释人类行为和意识的方方面面。但这类愿望总是会落空，因为构成心理学的不是一整套宏大的理论，而是许许多多不同的理论，每个理论仅仅能够解释行为的有限方面（Griggs, Proctor, Bujak-Johnson, 2002）。心理学的多样性使得理论整合变得极为困难。事实上，在许多心理学者看来，“整合”本身就是“不可能的任务”。尽管如此，另外一些心理学家却正在寻求领域内的理论整合（Cacioppo,

2007a, 2007b; Cleeremans, 2010; Gray, 2008; Henriques, 2011; Sternberg, 2005）。例如，在过去的20年间，心理学的学科统一性有所增强，这要归功于进化心理学家的努力。这些研究者将人类心理过程视为服务于某些重要进化功能（诸如亲缘关系识别、择偶、合作、社会交换及后代抚养等）的机制（Buss, 2005, 2011; Cartwright, 2008; Ellis Bjorklund, 2005; Geary, 2005, 2008），并试图以此来实现概念的整合。卡乔波（Cacioppo, 2007b）同样也指出，像社会认知神经科学这样的分支，将心理学中的大量专业如认知心理学、社会心理学和神经心理学紧密地联系在一起。

一些研究者认为心理学的多样性反映了学科的潜在优势（Cacioppo, 2007a; Gray, 2008）。例如，卡乔波（Cacioppo, 2007a）将心理学视为所谓的"枢纽学科"——会对其他领域产生广泛影响的学科。他引用证据表明，和其他科学相比，心理学的发现有着更深远的影响。然而，无论心理学家对心理学主题的统一性持何立场，他们都承认，即便有一天能够实现理论的整合，其过程也是极为困难的。缺乏理论的整合为心理学招来了一些批评，贬低了它作为科学所取得的进步。这类批评源于一个错误的观念，即所有真正的科学都必须具备一个宏大的、统一的理论。之所以说它错误，是因为它忽视了其他科学同样也缺乏一个完备统一的概念体系这个事实。哈佛大学心理学家威廉·艾斯特斯（William Estes, 1979）已经强调过这一点：

> 实验心理学家所面临的这种困境既不新鲜，也非独有。20世纪早期，物理学在本科水平的教学中便被分成了若干独立学科。因此，我是通过分别学习力学、热力学、光学、声学和电学而了解这门科学的。同样，化学也曾被分为无机化学、有机化学、物理化学和生物化学。当时这些分支之间的交流和融合并不比现在的心理学好到哪里去。它们都仅仅在抽象的数学理论水平上才实现了整合。医学也被分为众多分支，而且和心理学一样，没有新的整合出现（pp. 661-622）。

一旦我们知晓了决定某一门学科结构的社会和历史因素是怎么回事，就能认识到，要求所有领域具备高度统一性是不合逻辑的。事实上，许多学者认为"心理学"这个暗示学科内容具有统一性的词汇并不能反映学科的特性。其结果就是，许多美国的知名大学的学院已经把它们的名字改为心理科学学院（Department of Psychological Sciences）（见Jaffe, 2011）。"科学"这一术语承载了本章的两个重要信息：其一是多元化标志，即我们讨论过的这个学科内容的多样性的观点；其二"科学"这一词汇也标志着，应该从方法而非内容上去寻

找心理学这门学科的统一性。只有这样，我们才有望从不同研究者的研究目的中发现更多的统一性。但即使是在方法领域，也依然存在着有关这门学科的一些很深的误解。

科学的统一性

仅说心理学是关于人类行为的科学，并不能将它和其他学科区分开来。许多其他专业团体和学科——包括经济学家、小说家、法律、社会学、历史、政治科学、人类学和文学研究——都或多或少与人类行为有关，心理学在这方面并非独树一帜。

应用性也不能证明心理学具有任何独特性。例如，许多大学生选择主修心理学是因为他们有一个要帮助他人的崇高目标。但是在许多领域，如社会工作、教育、护理、职业咨询、物理治疗、警事科学、人力资源以及语言矫正等，“帮助他人”都是其中的重要组成部分。同样，培训应用性的、通过咨询来帮助他人的专业人才并不需要单独开辟一门叫做“心理学”的学科。

只有两点能证明心理学是一门独立的学科：其一，心理学研究采用科学方法来探究人类及动物的所有行为；其二，从这一知识衍生出的实际应用是具备科学基础的。如果不是这样，心理学就失去存在的理由了。

心理学不同于其他行为研究领域的地方在于，它试图向公众保证两点：第一，心理学中有关行为的结论都有科学证据；第二，心理学的应用都源于科学方法，并经过了科学方法的检验。心理学是否曾经偏离过这两个目标呢？有过，而且经常如此（Lilienfeld, 2007; Lilienfeld, Ruscio, Lynn, 2008）。本书就是关于怎样更好地实现这两个目标的。在第12章中，我将回到这一主题——一些心理学工作者因为不遵守适当的科学标准而自我损害了其作为心理学家的合法性。但是，从原则上讲，科学性正是保证心理学作为一门独立学科的标准。如果有朝一日心理学不再追求这些目标，即它不再愿意坚守科学标准，那它也就该关张大吉，将其关注的领域拱手让给先前提到的那些学科——因为此时它已成为了一个完全多余的知识领域。

毫无疑问，任何人想要理解心理学，第一步也是至关重要的一步，就是要意识到心理学的首要特征——它是有关行为的、以数据为基础的科学研究。对这一事实及其全部内涵的理解将贯穿本书的始终，因为这是我们认识真正心理学的最基本的途径。反过来说，人们

之所以对心理学的理解会出现各种各样的偏差，正是因为未能意识到它是一门科学的学问。例如，我们常常会听到学术圈外的人宣称心理学不是科学。为什么还会有这样的误解？

那些想让公众相信“心理学不能成为一门科学”的企图，其产生的背景各不相同。正如我们在后面的章节中所要讨论的，许多有关心理学的错误认识，都是由那些伪心理学的代理人处心积虑制造的。在我们的社会中，一个经营伪科学信念系统的巨大产业正在兴起，这一信念系统出于既得利益的考虑，总是想让大众相信，无论什么都能纳入心理学的范畴，而且心理学的主张不能以理性标准来衡量。这无疑为“催眠减肥”、“激发潜在心灵能量”、“睡觉时学法语”这类广告以及利润高达数百万的“心理自助”产业里其他诸多门道的营销创造了绝佳的氛围。此类门道要么不是建立在科学证据基础上，要么（在许多时候）与已有的证据相冲突。

另一种对于科学心理学的排斥是由于一些人不愿看到科学进入到长期以来由不容置疑的权威或“常识”统治的领域里。历史上此类例子不胜枚举——人们拒绝使用科学，而更喜欢利用哲学沉思、神学谕告或世俗智慧去解释现实世界。每一门科学都会经历过一个受到阻碍的阶段。与伽利略同时代的知识分子拒绝透过他的新望远镜观察天空，因为“木星存在卫星”颠覆了他们的哲学和神学信仰。几个世纪以来，人类解剖学的发展可谓步履蹒跚，因为世俗和宗教禁止对人类尸体进行解剖（基督徒认为，身体的内部是“上帝的辖区”，见Grice, 2001）。查尔斯·达尔文总是受到不断的抨击。保罗·布洛卡（Paul Broca）的人类学协会（Society of Anthropology）在19世纪的法国受到抵制，因为有人认为关于人类的知识会颠覆国家。

关于人类的知识每向前迈进一步，都会引发反抗。然而，当人们开始意识到科学并没有通过调查和研究对人性造成亵渎，而是以扩展知识的方式促进了人类的自我实现时，反抗终将烟消云散。谁现在还认为星系图以及宇宙是由无数星球所组成的复杂理论会摧毁我们对于宇宙的向往？谁会选择禁止人体解剖时的医疗保健系统，并进而拒绝从社区中获得的现代医疗保健？对于星球和人类身体的实证性态度并没有磨灭人性。更近的例子是，达尔文的进化论体系为遗传学和生物学取得非凡的进步奠定了基础。但是，在我们更接近人类的本质及起源的同时，残余的反抗势力仍然存在。在美国，一些政客继续施压，意欲在公立学校推行“神创论”教学；同时调查显示，有很大比例的美国人（欧洲和加拿大人亦如此）并不接受“人类是经过自然选择进化而来的”这一科学事实（Barnes, Keilholtz, Alberstadt, 2008; Frazier, 2009, 2010; Laden, 2008）。进化生物学有着无数辉煌的科学成就记

录，时至今日还是照样被公众所排斥。如此看来，心理学这门志在将所有关于人类的固有信念都置于科学检验之下的新兴科学，时下还会引发人们对其正确性的否定，这又有什么好奇怪的呢？

那么，什么是科学

为了理解什么是心理学，我们必须理解什么是科学。或许我们可以从“什么不是科学”入手。按这种方法，我们能摒弃大部分常见的错误观念。首先，科学并不是由内容来定义的。宇宙万物的任何方面对于一门科学学科的发展来说，都是一场公平的游戏，当然也包括人类行为的所有方面。我们不能将宇宙万物分为“科学的”和“非科学的”两类。尽管历史上始终有一股强大的力量，试图将人类排除在科学研究的范围之外，但正如我们所见，它们均以失败告终。拒绝将心理学作为一门科学学科来对待，可能代表了这一历史争论的余音。

科学也不能按照特定实验器材的使用来定义。试管、电脑、电子设备或研究者的白大褂都定义不了科学。这些都是科学的附属物而不是其本质特征。科学是一种思考和观察事物以便深入理解其运行机制的方法。

在本章的剩余部分，我们将讨论科学的三个相互关联的重要特征：（1）应用系统的实证主义；（2）产生公共知识；（3）验证可解决的问题。尽管我们将逐一检验每一条特征，但请记住这三条特征构成了相互联系的统一整体（更多有关科学的普遍特征的详细讨论，参见书后参考文献部分列出的Bronowski、Cournaud、Medawar、Popper、Raymo以及Sagan的著作）。

系统的实证主义

如果在任何辞典中查找“实证主义”，你会发现它的意思是“基于观察的实践”。科学家通过验证来找寻世界的规律。这个事实可能对你来说是显而易见的事实，而这正是过去两个世纪以来科学态度传播的结果。在过去，它却不是显而易见的。回想一下伽利略的例子，伽利略借助他那原始的望远镜，宣称看见了环绕木星的一些卫星。而在当时，有学识的人们都认为只存在七个“天体”（五颗行星、太阳、月球）。长久以来，人们认为获得知识的最佳途径是纯粹思考或诉诸权威。一些同时代的学者甚至拒绝透过伽利略的望远镜进行瞭望。还有

一些人认为望远镜是设计用来骗人的。更有一些人指责说望远镜是在地面上而非天空中工作（Shermer, 2011），另外一位叫弗朗西斯科·西奇（Francesco Sizi）的学者试图驳倒伽利略，但他并不是通过观察，而是通过下面的一番话：

> 人的脑袋上有七个窗口：两个鼻孔，两只耳朵，两只眼睛和一张嘴；因此在天界有两颗吉星，两颗灾星，两颗发光星（指日月），以及性状不明但无关紧要的水星。从这点和其他无数相似的自然现象诸如七种金属等中，我们就可以归纳出行星必然是七个……除此之外，犹太人和其他古老的民族，都将一周分为七天，并以七大行星来命名；如果现在我们增加了行星的数目，将导致整个系统的崩溃……进一步来说，卫星用肉眼无法看到，因此对于地球没有影响，既然没有用处，也就不存在（Holton & Roller, 1958, p. 160）。

关键问题不在于以上论述多么愚蠢可笑，而在于它被视作可与真实观察抗衡的一种辩驳。今天我们嘲笑它是因为我们都是事后诸葛亮。三个世纪以来，业已证明力量的实证取向使我们强于可怜的西奇。要是没有经历这些实证主义的岁月，我们中的许多人可能都会点头同意并对他大加褒奖。的确，实证取向并不一定显而易见，这就是为何即使在一个科学占统治地位的社会中，我们也不得不经常强调它的原因。

纯粹、单一的实证主义还不够。注意本小节的标题是“系统的实证主义”。观察很好，而且很有必要，但是对于自然世界单纯的、非结构化的观察并不能导致科学知识的产生。假使你记录下自己一天中从起床到睡觉之间观察到的所有情况，完成这一任务时，你会拥有一大堆事实，但仅凭此并不能让你更好地理解这个世界。科学观察被称为“系统性的”，是因为它是结构化的，所观察的结果能够揭示自然世界一些潜在的本质。科学观察通常都是理论驱动的；它们检验有关世界的各种不同解释观点。它们是结构化的，因此可以根据观察结果，决定哪些理论得到支持而哪些则被拒绝。

公共性的、可检验的知识：可重复性和同行评审

从某种特殊意义上说，科学知识是公共性的。当然，并不是说把科学发现张贴在社区中心的公告牌上就叫“公共性”了。我们指的是这样一个事实，即科学知识并不单独存在于特定个体的头脑之中。从某种重要意义上说，科学知识在没有提交给科学团体、接受他人批评和

验证之前，是根本不存在的。那些被认为隶属于特定个体思维过程中、不可接受他人审查和批评的“特殊”知识，永远都无法获得科学知识的地位。

科学通过可重复性来实现其公共可检验性的理念。一项发现如果想在科学界获得公认，就必须以一种能够让其他科学家尝试相同实验并获得相同结果的方式呈现给科学团体。当这一切都完成，我们就可以说，这一发现是可重复的。科学家利用可重复性来定义公共知识。可重复性保证了特定发现并不是由于个别调查者的错误或偏差而产生的。简而言之，一项发现如果想要被科学团体所接受，它必须能够被原始调查者以外的其他人所重复。当一项发现以这种方式被呈现出来时，它就成为了公共性的。它不再仅仅为原始研究者个人所有，它还能够被其他人获取、扩展、批评，或以他们的方式得到应用。

诗人约翰·唐尼（John Donne）告诉我们“任何人都不是一座孤岛”。在科学中，没有一个研究者是一座孤岛。每个研究者都与科学团体及其知识基础相联系。正是这种相互联系使得科学累积性地发展。研究者不断在原有知识的基础上进行新的探索，力求超越已知。而这一过程的前提便是，先前的知识以一种适当的方式予以陈述，使任何研究者都能以之为基础来进行探索。

公共性的、可检验的知识，指的是我们可以将研究发现递交给科学团体，团体中的任何人都能对其进行重复检验、批评或拓展。这个标准不仅对于科学家，同时对于作为消费者的外行人来说也是最重要的，因为他们必须对来自媒体的科学信息进行评估。正如我们将在第12章所看到的，区分大搞伪科学的江湖术士和真正的科学家最主要的一个方法就是，前者常常避开科学出版的常规渠道，而选择直接通过媒体公开他们的“发现”。当公众面对真实性可疑的科学发现时，一个屡试不爽、颠扑不破的标准就是，考察这些发现是否在得到认可的科学期刊上发表过以及是否经过了同行评审。对这一问题的回答往往能够区分“李鬼”和“李逵”。同行评审是指每一篇投到研究性期刊的文章都要经过数位科学家的评审，并将批评意见提交给编辑（通常都是此期刊所覆盖的某一领域中有资深研究经历的专家），再由编辑权衡这些意见，确定这篇文章可以立即发表，还是需进一步实验研究和统计分析之后再发表，或是因为有缺陷或价值太低而拒绝接受。大多数期刊在每期中都刊有编辑条例说明，因此很容易知道此期刊是否经过同行评审。

并非所有经过同行评审的科学期刊中的信息都必然正确，但至少它已满足了同行批评和监督的标准。同行评审是一个最低标准，而非严格的标准，因为大多数学科领域中都会有几十种质量参差不齐的期

刊。大部分科学观点在满足一些基本标准的前提下，都可以在某些地方以正规的方式出版。那种认为只有很小部分的数据和理论才能够在科学界获得出版的观点是错误的。当一些心理救助或治疗方面的江湖术士试图说服媒体和公众时，往往会暗示，一种所谓“正统科学”的阴谋将他们排除在科学出版渠道之外。但是，稍微想一想，心理学领域中有多少这样的合法渠道啊！APA的数据库 《心理学文摘》（PsycINFO）收录了来自2000种不同期刊的文献，其中大部分期刊都有同行评审。几乎所有待检验的理论和实验都能在如此众多的出版物中找到自己的发表渠道。

再次强调，我不认为所有发表在具有同行评审机制的期刊上的所有观点都必然正确。相反，正如我先前所强调的，发表只是一个最低的标准。然而关键在于，任何一种理念、一个理论、一项主张或疗法如果不能在有同行评审的学科文献中获得适当的收录，问题就很明显了。尤其是当某一主张缺乏证据却伴随着媒体的宣传活动时，此理念、理论或疗法显然是骗人的。例如，2005年美国宾夕法尼亚州有一桩著名的诉讼，有人试图在学校的生物课上教授神创论，鼓吹智力设计（神创论的一种形式）理念的一个证人说，他很难举出任何一个经过同行评审的有关智力设计的研究，尽管这一运动已经兴起了十年有余（Talbot, 2005, p.68）。

同行评审机制在不同学科之间有所区别，但是根本理念是相同的。同行评审是科学将客观性和公开评议标准制度化的一种方法（另一种是重复验证）。观点和实验要提交给其他评估，经过一个仔细推敲的过程。只有通过这一严格过程的观点才算符合了公共验证的标准。同行评审程序绝非完美，但它对我们消费者来说是唯一的保护机制。忽略它就等于让我们自己被巨大的伪科学产业玩弄于股掌之间，而这一产业又极其善于操纵媒体来达到自己的目的（见第12章）。在随后的章节中，我们将更详尽地讨论，如果忽视这些心理科学实践中固有的审查与制衡，我们将要付出多么高昂的代价。

实证可解的问题：科学家对于可检验理论的研究

科学针对的是可解决的、明确具体的问题。这意味着就其类型来说，科学家们所致力解决的问题是能通过现有的经验技术获得答案的。如果在当前所掌握的经验技术条件下，问题无解或理论不可验证，科学家们将不会对它展开研究。例如，“在日托期间，给予结构化语言刺激的3岁儿童，与那些没有给予这些额外刺激的儿童相比，是否可以更早地做好接受阅读指导的准备”就是一个科学问题，因为

在现有的经验方法之下，这是一个可解的问题。“人性本善还是本恶”就不是一个可实证的问题，因此不属于科学领域。“生命的意义是什么”同样也不是实证问题，因此也不属于科学领域。

科学通过以下方式得以进步：提出理论解释世界中的特定现象，根据这些理论作出预测，实证地检验这些假设，基于检验的结果对理论进行修正，通常次序为：理论→预测→检验→修正。因此对于科学家来说，“可解问题”这个词的意义通常是“可验证的理论”。什么样的理论才算是“可验证”的呢？这一理论必须与真实世界中可观察的事件具有特定的关联；这就是“实证可检验”的含义。可检验性标准在学术上通常被称为“可证伪标准”，这也是本书第2章的主题。

我们说科学家解决实证可解问题，并不是说在不同类别的问题中，有的本质上是可以解决的，而有的则注定无法解决，并且这种区分是固定不变的。恰恰相反，有些当前无法解决的问题，在理论和经验技术更加进步的时候会成为可解的。例如，20年前，对于“托马斯·杰斐逊是否与其奴隶萨丽·海明斯生下了她的某个孩子”这一争议话题，不会有历史学家认为它是一个实证可解的问题。然而到了1998年，由于基因技术的进步，这个问题已成为可解的，发表在《自然》（Nature）杂志的一篇文章（Foster et al., 1998）指出，杰斐逊极可能是埃斯顿·海明斯·杰斐逊的父亲。

这就是科学得以发展而新的科学得以诞生的方式。但对于“当前什么是可解的”这一问题，总是存在巨大分歧。因为涉及特定问题时，科学家们自己在这点上的意见都难以统一。因此，尽管所有科学家都认同可解性标准，但是他们对其特定应用可能存在不同的意见。诺贝尔文学奖得主彼得·密达沃（Peter Medawar）就曾把他的一本书定名为《可解的艺术》（The Art of the Soluble, 1967），并在书中指出，科学的一部分创造力就在于寻找处于人类知识最前沿、并可以用实证技术加以解决的问题。

心理学本身就提供了许多从无解到可解的好例子。有许多问题，诸如“一个孩子如何获得其父母的语言”、“为什么我们会忘记我们曾经知道的事情”、“身处一个群体中会如何改变一个人的行为和思想呢”等，在人们意识到可以以实证的方法来解答之前的几个世纪里，都只能被猜想而已。随着这一认识的慢慢发展，心理学逐渐集合了来自各个领域中关于行为的各种问题。心理学科逐渐脱离哲学，并成为了一门独立的实证科学。

认知心理学家史蒂芬·平克（Stephen Pinker, 1997）讨论了“未知”可以划分为“问题”或“玄谜”。如果是“问题”，我们知道其答案是能

找到的，即使我们目前还没有答案，我们也知道它大概是什么样子。如果是“玄谜”，我们甚至不能想象答案可能会是什么样子。利用这些术语，我们可以看到，科学就是将玄谜变为问题的过程。事实上，平克（1997）指出，他之所以要写《思维的运作》（How the Mind Works）这本书，正是因为“从心理表象到浪漫的爱情，几十个心理和思维方面的玄谜最近已经升级为问题了”（p. 9）。

心理学和世俗智慧：“常识”的误区

我们每个人都有一套内隐的有关行为的模型，这些模型影响着我们的互动以及我们如何看待自己和他人。事实上，一些社会、人格和认知心理学家正在探究这些内隐的心理学理论的本质。我们很少会清晰并有逻辑地表达我们的理论。相反，我们通常只有在特意关注它们，或者发现它们正遭遇某种挑战时，才会意识到其存在。其实，我们个人的行为模型并不像真正的理论那样具有内部一致性。相反，当我们觉得需要对行为作出解释时，往往搬出一箩筐关于人类行为的普遍真理、说教及谚语。这些关于行为的常识存在一个问题，它们之中有不少是自相矛盾的，因此也是不可证伪的（证伪原则是第2章的主题）。

人们爱用一些民间谚语来解释行为事件，即使之前在解释同一类型的事件时曾用过与之完全矛盾的谚语。例如，我们中的大多数人都听到或说过“三思而后行”——若不是我依稀记得之前有人告诫说“该出手时就出手”，我还会觉得这是个有用的、直接的行为建议呢！“小别胜新婚”明确预测了一种对于事件的情绪反应，但“眼不见，心不烦”不也同样如此吗？如果“欲速则不达”，那为什么有时我们又听到“时不我待”？既然“三个臭皮匠，顶个诸葛亮”，为什么又说“三个和尚没水吃”？如果我认为“行走江湖，安全第一”，为什么也相信“不入虎穴，焉得虎子”？如果“异性相吸”，为什么又“物以类聚”？我劝许多学生“今日事今日毕”，但我希望没跟我刚刚指导过的那个学生说过这番话，因为我刚还跟他说“要顺其自然”。这类谚语和俗话构成了对行为的固有“解释”，人们爱用它们，就是因为它们难以驳倒。不管发生什么事，可以拿一条出来解释一番。难怪我们都认为自己是判断他人行为和人格的高手。天底下发生的事我们都能解释。世俗智慧的可鄙之处就在于，它们压根不承担被驳倒的风险。

世俗智慧属于“后见之明”的智慧，并且它在真正预测性的意义上是无用的，这也就是为什么邓肯·沃茨（Duncan Watts）将他的一本书命名为《一切都显而易见：一旦你知道了答案》（Everything Is Obvious: Once You Know the Answer）（2011）。沃茨论述了拉扎斯菲尔德（Lazarsfeld）的一篇经典文章（1949），60多年之前，拉扎斯菲尔德在应对“社会科学不能告诉我们不知道的任何事情”这一普

遍的批评时，他列举了对600000名参加第二次世界大战的士兵进行调查后得到的一系列发现，例如，农村背景的士兵在服役期间比城市背景的士兵有更好的精神风貌。人们可能会倾向于认为所有的这些调查结果都显而易见。但是让我们以这个结果为例：人们倾向于认为，显而易见，农村出来的人更加适应严苛的物理环境，因此也肯定能更好地适应军旅生活。其他类似的结果同样显而易见。接着拉扎斯菲尔德使出了他的杀手锏：其实所有的发现都与最初呈现的陈述观点完全相反。例如，事实上，来自城市的士兵比来自农村的士兵在服役期间有更好的精神面貌。最后一部分的内容是为了让人们认识到，他们也能非常容易地对相反的结果作出解释。如果一开始就被告知了真实的结果，人们可能解释说他们认为城市的人已经习惯了拥挤的环境和等级化的权威。人们可能永远也不会意识到他们编造对相反的结果的解释是如此容易。

所以说，有时我们内隐的心理理论不容反驳。我们将在第2章中看到为什么这种不可反驳性造成了理论的失效。然而，即使我们的世俗观念有一些特定用处，甚至是经验可证的，也会产生问题。问题在于，心理学研究表明，在接受实证检验后，许多关于行为的普遍文化信念都被证明是错误的。

世俗观念（或称“常识”）出现谬误的例子俯拾皆是。比如说，有一种说法是，学习好或读书多的孩子都不擅长交际和体育。这个观点虽然错得离谱，但在当今社会上极为流行。有大量证据表明，与“常识”世俗观念正好相反，爱读书的人和追求学术成就者与不读书者相比，有着更强健的体魄，而且更常参与社交活动（Zill & Winglee, 1990）。再比如，学习成绩好的儿童比学习成绩差的更容易被同伴接纳。读书多的人比不读书者更愿意运动、慢跑、露营、远足、维修汽车等。

许多关于行为的世俗观念一经产生便生生不息。例如，20世纪90年代风行于社会和学校的一种世俗观念是，低自尊导致攻击行为。但实证研究显示，攻击行为和低自尊并无关联（Baumeister, Campbell, Krueger, & Vohs, 2003, 2005; Krueger, Vohs, & Baumerster, 2008）。相反，攻击行为似乎往往与高自尊相关同样。与之相似，过去20年间有一个非常流行的假说认为，低自尊会导致学业不良的问题。事实上，自尊和学业成绩之间的真实关系可能与教育工作者和家长的假设恰恰相反：是在校成绩（以及生活的其他方面）的优秀导致了高自尊，而非后者引起了前者。

考虑另一个常见的世俗观念——“孩子会给父母带来幸福”。如果我们考虑退休后孩子会带来的好处，这个陈述在一定程度上可能是对

的。人们回首往昔时，的确会发现孩子带来的幸福。问题在于，人们常常混淆回忆的观点和对真实事件的体验。从有孩子的例子看，两个观点是十分不同的。在年老时，觉得有个孩子确实令人开心。但是，就连续性、继时性的快乐（与追溯性的回忆相反）来说，孩子实际上让人的快乐减少。现在有一系列文献采用所谓“经验取样法”（experience-sample）观察人们在不同时刻的幸福程度（Brooks, 2008; Gilbert, 2006; Gorchoff, John, & Helson, 2008; Lyubomirsky & Boehm, 2010; Wargo, 2007），研究发现了一系列的趋势，例如结婚会增加幸福感。同时还发现，父母的幸福感会随着第一个孩子的降生降低。当第一个孩子成年时，幸福程度回弹一些，但随后又降低得更多。只有当最小的孩子离家自立后，婚姻幸福感才回归到没有孩子时的水平。

简言之，当世俗观念“孩子给父母带来幸福”接受科学检验时，情况就变得复杂起来。只有从回溯性的角度看，当孩子终于离开家时，我们可以品味将孩子抚养成人的成就感时，“孩子带来幸福”才是正确的。但这与这个世俗观念通常所表述的意思大相径庭。它通常指抚养孩子会让你立刻或在近期感到幸福。这就是世俗观念的惊人谬误之处。

世俗智慧是错误的另一个例子就是对学生来说常见的劝诫：做多选题时，即使对所选答案感到不确定，也千万不要更改最初的答案。不仅大多数学生认为他们不应该更改不确定题目的答案，而且《GRE巴朗指南》（Barron’s Guide to GRE）也建议“当你决定更改答案时要极其谨慎，经验表明更改答案的学生都会改错”（Kruger, Wirtz, & Miller, 2005, p.725）。这一建议完全是错的，错误的原因是，世俗迷思有关更改答案会降低一个人的分数的观点本身就大错特错。实际研究表明，当对一个多选题的答案有怀疑时，学生最好改变他们最初的选择（Kruger et al., 2005; Lilienfeld, Lynn, Ruscio, & Beyerstein, 2010）。

世俗智慧大行其道的例子还有关于“我们只用了10%的大脑功能”的世俗观念。虽然这一说法完全缺乏认知神经的基础（Boyd, 2008; Lilienfeld et al., 2010），但这一观念已流行了几十年，且俨然已成为所谓的“心理学公理”——即并不正确，但是已经重复了无数遍，以至于普通人认为就是事实的心理学观点。同样，一些人相信有些人是“左脑型”，有些人是“右脑型”；或者说人格的某些特定方面受左脑控制，另一些方面受右脑控制。尽管现代神经科学研究确实显示大脑中的一些功能专门化，但是对左右脑观点的这些通俗化说法却是胡说八道，尤其是在发现我们的大脑以联合的方式工作的情况下

（Lilienfeld et al., 2010; Radford, 2011）。

世俗观念并不总对证据免疫。有时，当与之矛盾的事实广为人知时，世俗心理学（“常识”）也会改变。例如，几年前，一个广为流传的有关儿童的俗语是“早熟者必早衰”（Fancher, 1985, p. 141）。这条俗语反映了“童年早熟与成年异常存在关联”这一信念，这一信念得到了许多“小时了了，大未必佳”的例子的支持。但在这件事上，心理学证据证明上述俗语并不准确，这一结论已被大众文化所吸收，所以你以后几乎不大会再听到这个世俗“智慧”了。

最后这个例子是一个警告，提醒我们注意当下的“常识”，因为不难看出，昨日的常识往往变成今天的谬论。毕竟，常识就是“尽人皆知的知识”，对吧？对。那么，人人都知道妇女不能投票，对吧？非裔美国人不应该接受教育，对吧？残障人士不该在社会里出现而应当被送到收容所去，对吧？事实上，150年前，这些观念都是尽人皆知的常识。当然，我们现在视这些过去的常识为谬论，都是些以完全未经证实的假设为基础的信念。但是，从这些例子中，我们可以看到心理学在常识面前扮演的关键角色。常识总是基于一些假设，而心理学对这些假设的经验基础进行检验。正如我们之前看到的许多例子，有时候假设得不到实证支持。这样的例子还有很多，通过它们，我们可以看到，心理学扮演着一种世俗智慧检验者的角色，常常难免和诸多根深蒂固的文化信念发生冲突。心理学往往是“传递坏消息的信使”[1]，宣告原本为人们所接受的世俗观念再无立足之地。这就不难理解，为什么许多人不仅无视这些消息，还想消灭这些信使。

注释

[1] 此处为“花剌子模信使”的典故，传说中亚古国花剌子模有位国王，会奖赏带来好消息的信使，而处死带来坏消息的信使。——译者注

心理学是一门年轻的科学

建立在实证基础上的心理学始终存在反对意见。仅仅100多年前，剑桥大学还拒绝建立一个心理物理学实验室，因为这样的一个主题研究，“以把人类的灵魂放在天平上的方式侮辱了宗教信仰”（Hearst, 1979, p. 7）。心理学致力于证明其问题是实证可解的，这一战斗也是最近才取得胜利。不过随着科学的进步，心理学家将涉足越来越多的主题，这些主题涉及人类某些牢固的信念，而很多都是可以通过实证方法验证的。心理学家现在正研究的话题性主题包括道德推理的发展、浪漫之爱、种族偏见的性质以及宗教信仰的心理和社会决定因素等。童年期性行为的研究最近引发了很多争议（Lilienfeld, 2010; Rind, 2008）。尽管有些人反对对这些领域进行实证调查，但这些领域都取得了科学进展。

美国心理学会前任主席杰拉尔德·库彻（Gerald Koocher, 2006）通过将他的一个首席专栏命名为“心理科学并不政治正确”来提醒我们有关心理学性质的问题。在文章中，他讨论了一些研究主题，例如肥胖的原因、决定政治态度的因素、宗教和性行为的关系、家庭暴力等。他指出这些标题下的研究结果都是有争议的，但是“心理科学不能被社会自由派或保守派制定的政治正确的标准所挟持”（p. 5）。

作为一门学科，心理学总是处于一种两难境地：一方面，一些人反对把心理学称为科学，否认心理学家可以建立关于行为的实证理论；另一方面，另一些人则由于惧怕心理学在某些行为领域揭示的真相会威胁到他们的信仰，而反对心理学家在这些领域进行研究。斯金纳学派的心理学家就总是面对这类相互矛盾的指责。例如，有批评者认为行为主义的强化法则不适用于人类行为。同时，另一些批评者则担心人们会运用这些规律去对人类进行严酷的、不人道的控制。因此，行为主义者腹背受敌，一些批评者否认行为主义者所发现的行为定律有用，而另一些批评者则害怕这些定律被滥用！

上述现象的产生主要是由于年轻的心理科学刚刚开始揭示行为方面的一些事实，而在过去，这些问题总是游离于研究之外的。它的青涩多多少少也解释了为什么许多人总是对这一学科产生误解。但无论如何，在过去的40年里，心理学已经在我们称之为科学的这个相互关

联的知识体系中站稳了脚跟。认识不到这一点，就会对心理学产生各种各样的误解。

小结

心理学是一个主题非常广泛，但又相对松散的学科，它包含一些通常不被归入同一概念的众多研究主题。然而，它们都使用科学方法来理解行为，从而实现了学科的统一。科学方法绝非是指一套生硬的规则，而是指一些非常普遍的原则。最重要的三点是：（1）科学采用系统的实证主义的研究方法；（2）它以可公开验证的知识为研究对象；（3）它研究实证可解的问题，并产生可检验的理论（第2章的主要内容）。构成系统实证主义基础的结构化及可控制的观察是本书随后几个章节的主题。科学通过同行评审等程序和重复验证等机制来保证知识的公共性。

心理学是一门新兴的科学，因而经常会和世俗智慧相冲突。这种冲突是任何新兴学科都会遇到的，了解这种冲突有助于我们理解为什么有人反对将心理学视为一门科学，并对心理学持敌意态度。同时，与世俗常识之间的碰撞，也令心理学成为一门激动人心的学科。很多人进入这一领域正是因为它提供了一个机会，让人们能够检验那些被毫无争议地接受了数百年的“常识”。

Chapter 2
可证伪性：如何挫败头脑中的小精灵

1793年，一场严重的流行病——黄热病袭击了费城。当时，这座城市里有一位顶尖的医生名叫本杰明·拉什（Benjamin Rush），他是美国《独立宣言》的签署人之一。在灾难过程中，拉什是少数几位确实治疗了几千例黄热病的医生。拉什信奉一种医学理论，认为黄热病必须用大量放血的方法治疗（用手术刀或水蛭吸血的方法使血液离开身体）。他为许多病人实施了这种疗法，当他自己感染这种疾病的时候，他也如法炮制。评论家指责他的治疗方法甚至比疾病本身更危险。然而，随着疾病的流行，拉什对他的疗法却更加自信了，即便曾有几个病人死去。这是为什么呢？

有人这么总结拉什的态度："一方面坚信自己的理论是正确的，另一方面又缺乏有效的方法对治疗效果进行系统研究，因此他将每个好转的病例都归为治疗方法的功效，而将每个死亡的病例都归为病情的严重性。"（Eisenberg, 1977, p. 1106）换句话说，如果病人情况好转，就被作为放血疗法有效的证据；如果病人死掉了，就被拉什解释为病人已经病入膏肓，无药可救。我们现在知道为什么对拉什的批评是正确的了：他的治疗方法和黄热病本身一样危险。在本章中，我们将要讨论拉什错在哪里。他的错误为阐明科学思维中最重要的一项原则提供了样本，而这一原则在评估心理学理论时尤其有用。

本章中，我们关注在第1章中已经讨论过的科学的第三个基本特征：科学只研究可解的问题。科学家们所说的"可解的问题"，通常是指可检验的理论。科学家要确认某个理论是不是可检验的，采取的方

法就是确保该理论是可证伪的，也就是说，理论对应着自然世界中的真实事件。接下来，我们就要看一看为何所谓的可证伪性标准在心理学中如此重要。

理论和可证伪性标准

本杰明·拉什在评估其疗法的效果时跌入了一个致命的陷阱。他的评价方法根本就不可能让人得出其治疗方法无效的结论。如果说，病人的恢复是对他治疗方法有效性的肯定（对其医疗理论的肯定），那只有当病人的死亡是对其治疗方法的否定时才算公平。但事实上，他却把这种否定合理化了。拉什解释证据的方式，违反了科学理论建构和检验应遵循的最重要原则之一：他令自己的理论不能被证伪。

科学理论的表述应该遵循这样的原则——从中得出的预测有可能被表明是错误的。因此，对某理论的新证据进行评价，必须使新的数据具有证伪该理论的可能性。这项原则通常被称为“可证伪性标准”。一位叫卡尔·波普尔（Karl Popper）的哲学家一直致力于强调可证伪性标准在科学进程中的重要作用，现在他的文章仍被从事科研工作的科学家们广泛阅读。

可证伪性标准主张，一项理论如果有用，它所作出的预测必须是明确的，理论必须两面兼顾，也可以说，这项理论在告诉我们哪些事情会发生的同时，应该指出哪些事情不会发生。如果不会发生的事情确实发生了，我们就得到了一个明确的信号——这项理论有问题：它可能需要修正，或者我们需要去寻找一个全新的理论。不管哪种方式，我们将最终有一个更接近真理的理论。相反，如果一项理论预测包括了所有可能观察到的数据，那么它将永远不能被修正，同时我们将被禁锢在当前的思维方式中，失去了取得进步的可能。这就是说，一项成功的理论并不是可以用来解释所有可能的结果，因为这样的理论本身就丧失了任何预测能力。

在这本书的余下部分，我们会经常涉及理论的评估，因此我们必须澄清一个关于理论的常见误解。这个误解体现为我们常说的一句话：“哦，这只不过是一种理论。”这句话代表了外行人使用“理论”这个词时通常所指的意思：一项未经证实的假设，一个纯粹的猜想或直觉。这意味着一个理论与其他理论并无优劣之分。“理论”这个词在科学上绝对不是这么用的。当科学家说到“理论”的时候，他们指的不是未经验证的猜想。

科学上的理论是一组具有内在联系的概念，它们能对一组数据作出解释，并对未来实验的结果作出预测。假设是从理论中产生的具体预测（理论则更加普遍和全面）。目前可行的理论是那些产生了一些假设，并且其中许多已经得到了验证的理论。因此这种理论的理论结构与大量的实证观察相一致。然而，当观察数据开始与理论中提出的假说相矛盾的时候，科学家们会尝试构建一个能为数据提供更好解释的新理论（或者在更通常的情况下，只是修正已有的理论）。因此，目前在科学范畴内所讨论的，都是在一定程度上已经被证实了的、所作出的预测并没有与现有的数据相矛盾的理论。它们并非纯粹的猜想和直觉。

外行人和科学家们使用“理论”这个词时的这种差异，经常会被一些试图将神创论纳入公立学校教育的虔诚的正统基督教徒所利用（Miller, 2008; Scott, 2005）。他们的论点通常是“进化论毕竟只是理论”。这种观点试图借用外行人对“理论”术语的用法，蓄意将理论歪曲为“只是一个猜想”。然而，通过自然选择的进化理论不是外行人所理解的“理论”（相反，在外行人的理解中，它应被称之为“事实”，见Randall, 2005），而是一个科学意义上的理论，是由一系列庞大而多样的数据支持的结论（Dawkins, 2010; Shermer, 2006; Wilson, 2007）。它并不等同于其他任何猜想，不是一个纯粹的猜测。相反，它与从属于其他学科的知识密切相关，这些学科包括地质学、物理学、化学以及生物学的各个分支。著名的生物学家西欧都萨斯·杜赞斯基（Theodosius Dobzhansky, 1973）在他的一篇题为《生物学中除了进化论以外，别无他物》（Nothing in Biology Makes Sense Except in the Light of Evolution）的著名文章里就阐述了这一观点。

敲门节奏理论

下面假设的例子展示了可证伪性标准是如何起作用的。一个学生在敲我的门。跟我同一办公室的同事有一套“不同的人以不同的节奏敲门”的理论。在我开门之前，我的同事预言门后是一位女性。我打开门，这个学生确实是女性。事后我告诉同事，他的表现令我惊叹，但这种惊叹程度非常有限，因为即使没有他所谓的“敲门节奏理论”，他也有50%的正确几率——实际上这一概率要更高一些，因为在大多数校园里，女性都比男性多。他说他的预测能高于随机水平。另一个人来敲门，我的同事预测说，这是个男性，而且不到22岁。我打开门，果然是个男生，而且我知道他刚从中学毕业。我承认我有点被震撼了，因为我所在的大学有相当数量的学生是大于22岁的。当然，我仍然坚持说，校园里年轻的男性相当普遍。见我如此难以被说服，我

的同事提出做最后一次测试。在下一个人敲门之后，我的同事预测：女性，30岁，5英尺[1] 2英寸[2] 高，左手拿书和挎包，用右手敲的门。打开门后，事实完全证明了预测，对此我的反应截然不同了。我不得不说，如果我的同事不是使用诡计事先安排这些人出现在我门口的话，我现在的确非常震惊。

为什么我的反应会不同呢？为什么我同事的三次预言会让我产生三种不同的从“那又怎么样”到“太不可思议了”的反应？答案与预测的具体性和精确性有关。越精确的预测在被证实的时候会给我们越大的触动。要注意，不管怎样，精确性的变化和可证伪性直接关联。预测越具体和精确，有可能证伪它的观测现象就越多。例如，有很多不是30岁和5英尺2英寸高的女性。请注意这里暗含的信息：从我截然不同的反应可以看出，一个能够最大限度地将不可能发生的事件预测出来的理论最容易令我折服。

好的理论作出的预测总是会显示自己是可证伪的。坏的理论不会以这种方式把自己置于危险的境地，它们作出的预测是如此笼统，以至于总会被证明是正确的（例如，下一个来敲我门的人会是100岁以下），或者，这些预测会采用一种能免于被证伪的措辞方式（如本杰明·拉什的例子）。事实上，当一种理论被置于“不可被证伪”的保护下，那么可以说它已经不再是科学了。事实上，哲学家卡尔·波普尔正是由于试图界定科学和非科学的区分标准，才会如此强调证伪原则的重要性。这里的讨论和第1章中我们有关弗洛伊德的讨论，甚至与心理学之间都有直接的联系。

弗洛伊德与可证伪性

在20世纪最初的几十年，波普尔一直在探寻，为何一些科学理论似乎导致知识的进步，而其他一些则导致智力停滞（Hacohen, 2000）。例如，爱因斯坦的广义相对论引发了一系列惊人的发现（例如，从一个遥远的恒星发出的光线经过太阳附近时发生弯曲），恰恰是因为它是这样建构预测的：许多事件或现象一旦被证实与之相矛盾，就可以证伪该理论。

波普尔指出，一些使知识停滞的理论却并非如此，并以弗洛伊德的精神分析法作为例子。弗洛伊德的理论使用一个复杂的概念结构，在事后解释人类行为，但并不作事前的预测。简而言之，弗洛伊德的理论可以解释一切，但是波普尔认为，也正是这个属性使得它在科学上毫无用处。它不作具体的预测。精神分析理论的拥护者花费大量的

时间和精力试图用他们的理论解释人类所有已知的活动——从个人的怪癖行为到广泛的社会现象，但他们在使这个理论成功地成为事后解释的丰厚资源时，也剥夺了其所有的科学实用性。如今，弗洛伊德的精神分析理论在激发文学想象力方面的作用比在当代心理学中扮演的角色更重要。它在心理学中的地位日益下滑，部分原因就是未能满足可证伪性标准（Wade & Tavris, 2008）。

这种不可证伪理论的存在会导致实际的危害。例如，对于孤独症（部分基因紊乱所致的疾病）成因的解释就曾被精神分析的解释带入了死胡同。受到精神分析学派的影响，心理学家布鲁诺·贝特海姆（Bruno Bettelheim）让“冰柜母亲”这个如今已被证伪的概念在当时广为流传，认为“造成婴儿时期孤独症的原因是父母不希望孩子存在”（Offit, 2008, p.3）。像这样的观念不仅有破坏作用，还阻碍了孤独症的研究。

作为另一个例子，回想一下抽动性秽语症的历史。这是一种以身体抽搐和痉挛为特征的紊乱，并伴有言语症状，如嘟囔、吠叫、模仿言语（无意识地重复他人的话）和秽语（强迫性重复淫秽词语）。抽动性秽语症是一种器质性的中枢神经系统紊乱，并已经成功地被药物治疗所攻克（Scahill et al., 2006; Simth, Polloway, Patton, & Dowdy, 2008）。纵观历史，抽动性秽语症患者一直遭受着迫害，早期被宗教统治者视为妖魔，近代又被认为是鬼怪附体，要被强制驱魔（Hines, 2003）。更重要的是，在1921至1955年之间，对这种病的解释及疗法一直被精神分析学派的概念体系所把持，这在很大程度上阻碍了人们对此病成因及治疗的理解（见Kushner, 1999）。有关这种病症的不可证伪的精神分析解释层出不穷。这些似是而非的解释所造就的概念泥潭蒙蔽了这一病症的实质，也阻碍了对其进一步的科学探究。例如，有一位作者曾经这样写道：

> （抽动性秽语症是）精神分析导致脑部疾病研究发生倒退的典型例子。勒·图雷特（La Tourette）将疾病归因于大脑的退行性变化过程。而在20世纪最初的几十年，由于弗洛伊德理论的盛行，对这种病的关注偏离了大脑……这一倒退的结果使病人往往被转到精神科医生（通常是精神分析学派的医生）而非神经科医生那里，因此没有接受生理检查和研究（Thornton, 1986, p. 210）。

夏皮罗等人（Shapiro, Bruun, & Sweet, 1978）提到，一位精神分析师认为，他的病人“不愿意放弃抽动，因为这成了她性快感的源泉和潜意识性欲的表达”；另一位精神分析师则认为抽搐“等同于手淫……与生殖器快感相联系的力大多转移到了身体的其他部位”；第

三位认为抽搐是一种“肛门施虐的迁移症状”；第四位认为，抽动性秽语症的患者具有“强迫型人格以及自恋倾向”，病人的抽动“代表了一种情感症状，对想表达情感的压抑性防御”。

事实上，这类例子不胜枚举，在他们无知自大的过度自信中尤为典型。发展心理学家杰罗姆·卡根（Jerome Kagan, 2006）告诉我们：“弗洛伊德的弟子桑德尔·费伦齐（Sandor Ferenczi），其本人从未见过抽动性秽语症病人，但是却犯下了同样严重的错误，他曾写过，抽动性秽语症病人的频率性的面部抽搐是由于对手淫的压抑。”（p.179）

夏皮罗等人（1978）对这类理论现状的总结，很好地说明了忽视可证伪性标准的有害影响：

> 精神分析这种理论化的方式简直面面俱到。抽搐是迁移性的症状而非歇斯底里症、肛门的而又是性欲的、受意志控制的而又是强迫性的、器质性病变同时又与原始心理动力有关……这些心理标签、诊断和治疗被不幸而且是以一种毫不谦卑、相当武断、伤害巨大的方式强加在病人及其家属身上。因为其随后的广泛影响，这些观点为对此病症的认识和诊治造成了极大的障碍（pp. 39-42, 50, 63）。

当研究人员承认精神分析的“解释”对治疗该疾病毫无用处的时候，对抽动性秽语症的认识和治疗才开始获得进展。那些毫无用处的解释是诱人的，因为它们似乎能对事情进行解释。事实上，它们都是在事后对所有事情作出解释的。然而，它们提供的解释不过是制造了理解的幻觉。由于总试图在事后解释一切，它们也就堵死了前进的大门。只有当一种理论并不预测所有事情，而是提出具体的预测——提前告诉我们哪个特定的情形会出现时，该理论才会出现进步。当然，从这样的理论推导出的预测可能是错误的，但这是优势，而非缺点。

小精灵

如果人们能够从所研究的问题里跳出来，尤其是人们若能以史为鉴（如本杰明·拉什的例子）的话，就不难识别出那些不能证伪的概念体系。当其例证明显是编造的时候，也很容易察觉其不可证伪性。举例来说，大家还不知道，我已经发现有一种大脑机制在控制行为，你很快就会在随处可见的八卦杂志上看到这个发现。我发现在大脑左半球的语言区附近住着两个小精灵，它们有能力控制大脑许多区域中的

电化学过程。而且，长话短说，它们基本上控制了一切事情。但是，有一个问题阻止我们看到它们，那就是小精灵有能力发现任何对大脑的侵入（外科手术、X光等），一旦觉察到外界的探测，它们就会消失（我忘记说了，它们具备隐身能力）。

毫无疑问，我在这里是用一个更适合小学生的例子来侮辱你的智慧。很明显，这个例子是我捏造的，但我对小精灵的假设永远无法被证实是错误的。不妨考虑一下。作为心理学导论的讲师和公开演讲者，我经常被问到，为什么不讲授在过去几年里在超感官知觉（extrasensong perception, ESP）和通灵学方面取得的那些惊人的新发现。我不得不告诉这些提问者，他们所获悉的大多数相关信息，无疑都是来自于大众媒体，而非科学界所承认的信息来源。事实上，一些科学家曾关注过这类说法，但没能够重复这些发现。我要提醒各位读者，要将一个研究成果认定为确定的科学事实，可重复性是至关重要的，尤其是当研究结果与以前的数据或现有的理论相矛盾的时候。

我甚至可以坦率地说，许多科学家对ESP研究已经失去了耐心。原因当然与此领域充斥着欺诈、江湖骗术和媒体炒作的现象有关，但令科学界觉醒的更重要的原因是马丁·加德纳（Martin Gardner, 1972）所谓的“ESP研究的22条军规”。其运作方式如下：一名“信奉者”（在开始调查之前就相信ESP现象存在的人）声称已在实验室证明了ESP。一名“怀疑者”（质疑ESP存在的人）被邀请证实这种现象。通常，在观察实验情境之后，怀疑者会要求信奉者进行更多的控制（我们会在第6章中讨论这种类型的控制），虽然这些要求有时候会被拒绝，但通常善意的信奉者们会同意他们的要求。当加入了实验控制之后，这种现象就不再出现了（Farha, 2007; Kelly, 2005; Milton & Wiseman, 1999; Park, 2008;Wiseman, 2011）。怀疑者会对这种失败作出正确的解释——早先对这个现象的证实是由于缺乏足够的实验控制，因此结论不能被接受。但他们往往吃惊地发现，信奉者并不承认早先的证明是无效的。相反，他们搬出超感知的“22条军规”：他们坚称，心理能量是很敏感的、微妙的，并极易受到影响。怀疑者的“负面感应”是瓦解这一“超感官能量”的罪魁祸首。信奉者认为，怀疑者的“负面气场”被移开后，这种心理能量无疑将会回归。

这种对无法在实验室中证实ESP的解释方式，在逻辑上与我编造的小精灵的故事相似。ESP的运作就像小精灵一样。只要你不侵入性地仔细观察它，它就在那儿。如果你观察它了，它就不见了。如果我们接受这种解释，那么向怀疑者证明这一现象就变得不再可能。这种现象只为信奉者现身。当然，这种说法在科学领域是不能接受的。我们没有磁力物理学家和非磁力物理学家之分（即磁场只对前者存

在）。以这种方式解释ESP的实验，使得ESP的假设变得像小精灵的假设一样不可证伪。正是这种解释方式，将ESP排除在了科学殿堂之外。

不是所有的证实都等价

可证伪性原则对于我们如何看待一个理论的证实过程具有重要的意义。许多人认为，一个好的科学理论就是被多次证实的理论。他们假设，被实的次数是对理论进行评价的关键。但是，可证伪性原则意指理论被证实的次数并不是最重要的因素。原因在于，正如“敲门节奏理论”所展示的那样，并不是所有的证实都是等价的。证实能否令人信服，取决于预测在何种程度上将自己暴露在可能被证伪的情境下。一个非常具体的、可能被证伪的预测（例如，一位女士，30岁，5英尺2英寸高，左手拿书和挎包，用右手敲门），比20个不可证伪的预测（例如，一个小于100岁的人）拥有更强的说服力。

因此，我们不能仅关注理论被证实的数量，更要关注验证本身的质量。将可证伪性作为一种评价标准，就可以使那些使用研究结果的人抵制不科学的、全能理论的诱惑。这种全能理论会不可避免地妨碍我们对世界和人类本质进行更深入的探索。事实上，这种理论的死角也正是最魅惑人的地方，因为它们永远不能被证伪。在纷繁多变的现代世界中，这种理论千年不变。

波普尔经常指出：“这些（不可证伪的）理论拥有巨大的心理吸引力，其秘密在于它们能够解释一切事情。预先知道无论什么事情发生，你都能理解它，不仅给你智力上的掌控感，而且，更重要的是，让你拥有应对这个世界所需的安全感。”（Magee, 1985, p. 43）但是，这种安全感的获得并不是科学的目标，因为对这种安全感的追求是以知识发展的停滞为代价的。科学是一套不断挑战原有信念的机制，在这种机制里，原有信念以一种能够被证伪的方式接受实证检验。这一特点往往使科学（尤其是心理学）与所谓的世俗智慧或者常识直接发生冲突（正如我们在第1章中所讨论的）。

可证伪性和世俗智慧

心理学威胁到世俗智慧所提供的安逸感，因为作为一门科学，它不能只提供无法被反驳的解释。心理学的目标是对各种行为理论逐一进行实证检验和筛选。某些世俗智慧表述得很清晰，经得起实证检

验，这当然是心理学所欢迎的，而且其中许多已经被纳入了心理学理论。然而，心理学并不追求那类事后能解释一切，但事先无法作出任何预测的理论，不追求这种解释系统所带来的安逸感。它不接受那些被设计得永不可变、并代代相传的世俗智慧体系。试图向学生和公众隐瞒这一点无疑是自毁长城。不幸的是，一些心理学指导教师和普及者觉察到了心理学对世俗智慧的威胁给一些人造成的困扰，于是他们有时会试图通过传递错误信息来安抚这种情绪，如“你会学到一些有趣的东西，但别担心，心理学不会挑战那些你深信不疑的观点”。这是一个错误，并且会对“什么是科学”和“什么是心理学”造成混乱。心理学建构了有关于性行为、智力、犯罪、经济行为及其他诸多人们感受强烈的主题。如果对上述这些主题的调查没有发现一些令某些人感到颠覆的事情，这才奇怪呢！

科学寻求概念上的变化。科学家试图描绘世界的真实图景，而此图景可能与我们的固有信念正好相反。现代思潮中有一种危险的趋向——认为应避免让一般大众知道世界的真正本质，一种无知的面纱是必要的，以防公众面对真相时手足无措。心理学与其他科学一样，拒绝向人类隐瞒真相。

此外，当我们被那些对人类行为有误解的人们所包围的时候，大家都会蒙受损失。公众对于教育、犯罪、健康、生产力、儿童福利和许多其他重要问题的态度塑造了我们的世界。如果这些态度源于错误的行为理论，那么我们大家都会受到伤害。

承认错误的自由

科学家们发现，可证伪性原则的一个最具解放意义和最有用的启示是：在科学上，犯错并不是罪过。哲学家丹尼尔·丹尼特（Daniel Dennett, 1995）曾说过，科学的本质就是“在公众面前犯错——在众目睽睽下犯错，以期他人能够帮助其修正这些错误”（p. 380）。当数据与理论不符时，通过对理论进行不断地修正，科学家们最终共同构建起能更好地反映世界本质的理论。

事实上，如果我们能够在日常生活中使用可证伪性原则的话，我们的生活品质将会大大改善。这就是为什么我在本节的第一句话中使用“具有解放意义”这个词的原因。它包含着一种个人化的期许，即此处产生的理念能够同时对科学之外的领域有所启示。如果我们能够理解这一切，当我们的信仰与观察到的事实相冲突时，我们最好是调整信仰，而不是否认事实和坚持错误的想法，这样我们将会很少遇到一

些个人和社会问题。

当你与某人激烈地争论的时候——也许就是当你给出一个有力的反击来捍卫你的观点的时候——有多少次你会突然意识到你搞错了某个关键事实或论据？这时你会怎么做？你会收回前面的话，并向别人承认错误，同时承认别人的解释现在看起来比你的更合理吗？或许不会。如果你和我们中的大多数人一样，那么你一定会采用“无限合理化”的策略，即没完没了地寻找一些理由为自己先前的错误辩解。你试图在不承认失败的情况下使自己从争论中全身而退。你最不可能做的就是承认自己错了。这样的话，你和争论对手都会更加疑惑：到底哪一种信念更接近真理？如果争论不能成为公共性的（如在科学中那样），如果正确和错误的信念以同样激烈的方式争辩，如果争论的结果不能得到正确的反馈（如本例），那就没有更可靠的机制使信念与现实相吻合了。这就是为什么那么多私人和公开的对话令人困惑，为什么相比所谓的常识或世俗智慧，心理科学在解释人类行为的原因方面更加可靠。

如果理解了我们的信念在历史上多么地因事而异，可能就会发现我们在证据面前是多么容易改变自己的信念。也就是说，信念的改变与否仅仅取决于他们成长时在某地某时发生的事件的多少。然而调查显示，大多数人对这种情况不够重视。在我自己的实验室中，我们对设计来评定人们鉴别信念的历史可变性的能力问卷中的数据进行分析。以问卷中的一个项目为例，被试必须对于“即使我周围环境（家庭、邻居、学校）与现实情况不同，我也会有与现在相同的宗教思想”这一陈述表达“非常、中等地或些许地同意或者不同意”。宗教问题是信念随环境变化的典型案例（欧洲、美国人信奉基督教；非洲和中东人信奉伊斯兰教；印度信奉印度教等）。虽然如此，我和我的同事在几个调查中发现，大约40%到55%的大学生会认为他们的宗教思想无论如何也不会随着他们的历史环境（父母、国家、教育）而改变。

证伪态度有益于科学本身的原因是，科学通过排除不正确的假设而非集中精力研究完美的理论而进步，这种态度在研究一个问题的初期尤为重要。事实上，生活中诸多领域的情况都是如此。通常定位最正确的表现方式是很困难的，但是聚焦表现的错误却相对容易。评论家尼尔·波兹曼（Neil Postman, 1988）指出，医生很难去定义“完全健康”，但却很擅长发现疾病。同样，律师能够轻易判断不公正，却很难定义“完全公正”。证伪态度对于科学家来说很有用也正是基于这一原因。尤其在研究问题的早期，研究的焦点在什么是错误的——排除不正确的信念——对科学家来说是很有成效的方法。

在科学过程中犯错是正常的，对于科学进步来说，最大的危险在

于我们人类极力避免将自己的固有信念暴露在可能被证明是错误的环境中的倾向，许多科学家已经意识到了这一观点的重要性。科学家们必须避免这种倾向，诺贝尔奖得主彼得·米德瓦（Peter Medawar, 1979）认为科学家应当牢记：一个假设在何种程度上被确信为正确，实际上与其是否为真无关（p. 39; 原文为意大利语）。

可以通过这样一种途径来理解米德瓦的话。喜剧演员斯蒂芬·科尔伯特（Stephen Colbert）在其2005年12月17号的表演中，创造了术语“感实性”。感实性是指“内心感到真实，但是没有证据支持”（Manjoo, 2008, p.189）。米德瓦说的意思是，科学拒绝感实性。这常常使科学在现代社会中成为异类，因为现代社会中，感实性比以往更为普遍。

心理学界许多最具声望的科学家都遵循米德瓦的建议——“一个假设在何种程度上被确信为正确，实际上与其是否为真无关。”在一篇报道实验心理学家罗伯特·克诺德（Robert Crowder）职业生涯的文章中，引述了他的一位同事马扎林·巴纳吉（Mahzarin Banaji）的话：“他是我认识的最不维护自己理论的科学家。如果你发现一种方法证明他的理论有漏洞，或者他的实验发现有局限性或有缺陷，他会非常高兴，并会和你一起计划如何推翻该理论。”（Azar, 1999, p. 18）艾泽（Azar, 1999）描述了克诺德如何提出了一个叫作“前分类听觉存储器”的记忆成分理论，然后又仔细地设计了一个实验研究证伪了自己的模型。物理学家杰罗米·格鲁普曼（Jerome Groopman, 2009）描述了证伪态度在医疗诊断中是多么实用和有效：“所以医生学会去质疑他提取的病历数据的有效性……最有启发性的时刻是你被证实是错误的，并且意识到你相信自己知道的比实际的多，其实是忽略了能够否认你的推诊或是未能考虑到病人患有不止一种疾病的关键信息。”（p.26）

早于达尔文数百年前，亚里士多德曾经说过：“受过教育的标志是可以去思考一种思想，而不去接受它。”更有意思的是，经济学家约翰·梅纳德·凯恩斯在经济大萧条期间阐明了证伪态度，他批判亚里士多德道：“事实改变，我就会改变我的观念。你能吗，先生？”（Malabre, 1994, p.220）

但是，要让科学发挥作用，并不需要每位从事科学工作的科学家都具备证伪的态度。科学那种揭示世界真知的独特力量，并不产生于科学家们独特的德行（即他们是完全客观的，他们在解释研究结果时从来不带偏见等）。实际上，这种力量的产生是因为会犯错的科学家们身处一个证实与平衡的程序中。在这个程序中，总会有其他科学家提出批评并发现他们同行的错误。哲学家丹尼尔·丹尼特提出过相同的

论点："不是每位科学家都必须表现出罗伯特·克诺德的客观性。"丹尼特强调"科学家和其他任何人一样容易犯错，但认识到他们及其所属团体的犯错根源之后，他们设计出精巧的系统来约束自己，努力防止自身弱点和偏见影响自己的研究结果。"（p. 42）心理学家雷·尼克尔森（Ray Nickerson, 1998）以一种更为幽默的方式道出了相同的观点："科学家们的虚荣心实际上在科学进程中起着作用，"科学家对自己的想法抱有的批判性态度并没有在很大程度上导致科学的成功……更真实的情况是，每个科学家都积极地想要证明某些科学家所持有的观点是错误的。"（p. 42）科学知识的力量并不是来自于科学家的德行，而是源于他们不断交叉检验彼此的知识和结论的这一社会过程。

想法不值钱

先前关于检验世俗智慧的讨论，将我们引向了可证伪性原则的另一个有趣推论：想法不值钱。说得更准确些，我们的意思是某些类别的想法不值钱。生物学家和科学作家史蒂芬·古尔德（Stephen J. Gould, 1987）对此有所阐述：

> 15年的月刊专栏写作生涯，让我收到各个科学领域非专业读者的海量来信……我发现一个常见的、同时是压倒性的错误观点。人们会告诉我他们提出了一项革命性的理论，它会拓展科学的边界。这些理论通常以单倍行距打印在几张纸上，内容通常是对最深层的终极问题的猜测——什么是生命的本质？宇宙的起源？时间的起点？但是，这些想法不值钱。任何智力正常的人都能在早饭前想出几个这样的念头。科学家们自己也很容易就能想出来。但我们不这样做（或者说，我们只让它们留在自己脑子里），因为我们不能找到方法来验证它们以决定它们的对错。一个既不能被证实也不能被证伪的可爱想法，对科学来说又有什么用呢？

古尔德对最后一个问题的回答是："没用。"古尔德这里所说的廉价想法正是我们早先在对卡尔·波普尔观点的讨论中提到的那些：包罗万象、复杂、"模糊"、能够用来解释一切的宏大理论——这种理论的建构更多是为了提供情感支持，因为它们没打算被改变或抛弃。古尔德告诉我们，这种理论对于科学目标是无用的，无论它们多么有抚慰功能。科学是创造性的过程，但是这种创造性需要让概念结构符合实验数据。这并不容易做到。那些如实解释真实世界的想法一点儿也不廉价。也许这就是为什么好的科学理论很难提出、而不可证伪的伪科学信仰体系泛滥的原因，因为后者很容易建构。

科学理论与世界紧密联系。它们是可证伪的，并能作出明确具体的预测。事实上，形成真实的、科学真正可以解释的理论是一项困难的任务。但是，理解科学运作的一般逻辑并没有那么困难。事实上，现在已经出版了大量专为儿童撰写的关于科学思维逻辑的书籍（Binga, 2009; Bower, 2009; Dawkings, 2012; Epstein, 2008; Swanson, 2001, 2004）。

注释

[1] 1英尺≈0.305米。——译者注

[2] 1英寸 = 2.54厘米。——译者注

科学中的错误：逼近真理

在解释可证伪性原则的过程中，我们已经勾勒出科学进步的简单模式：提出理论、从中推导出假设，然后让假设接受各种技术或方法的检验——我们将在本书余下的部分讨论这些技术。如果假设通过了某些实验的检验，该理论就得到了某种程度的确证；如果假设被实验证伪，这个理论就得做出某种程度的改变，或者被一个新理论所取代。

当然，虽然科学上的知识是暂时性的，由理论得出的假设可能是错误的，但这并不是说所有的一切都要被拿来检验一番。科学中有很多理论已经被确认过无数次，它们被称为“公理”，因为它们几乎不可能被未来的实验推翻。我们不大可能在某一天发现，血液不是循环的，或者地球并没有在环日轨道上。这些众所周知的事实并不是我们一直在讨论的假说。它们也不是科学家们的兴趣关注点，因为它们已经是确定无疑的。科学家只对已有知识范围之外的问题感兴趣。对于确定无疑的事实，他们毫无兴趣。

科学实践的这一面——科学家侧重于已知事实的前沿，而忽视那些已经被充分证实的问题（所谓的公理）——对大众来说很难理解。科学家们似乎总是更强调未知的事物而非已知事物。这千真万确，而且科学家有很好的理由这么做。为了推进知识的进步，科学家们必须一直身处已知的前沿。当然，这里是很多事情都不确定的地方。但科学进步正是通过这个过程来实现的，即试图在已知的前沿减少不确定性。这种特点常常使得科学家被公众视为“没谱的”。但这只是表面现象，科学家们只是对知识的前沿不确定——这使我们对于事物的理解不断加深。科学家们不怀疑那些被很多研究重复证实的事实。

同样需要强调的是，当科学家通过观察法证伪一个理论或用一个新理论代替旧理论的时候，并不意味着他们要将先前用以建立旧理论的事实全都扔到一边（我们会在第8章展开讨论这个话题）。相反，新理论应该能够解释所有旧理论能解释的事实，还能够解释旧理论不能解释的事实。理论被证伪并不意味着科学家非得建构一个全新的理论。复杂的理论通常是大体正确而非完全正确的，大体正确的判断可能更接近事实的真相（Radcliffe Richard, 2000）。

科普作家伊萨克·阿西莫夫（Isaac Asimov）在一篇题为《错误的相对性》（The Relativity of Wrong, 1989）的文章中很好地说明了理论修正的过程，文中谈到我们对地球形状的理解是如何完善的。他首先提醒我们，不要以为“地球是平的”这一古老信念是愚蠢的，在平原上（大部分有文字的人类文明都发源于平原），地球看上去相当平坦。阿西莫夫要求我们试着对不同的理论进行定量的比较，看结果会告诉我们什么。首先，我们能够将不同理论表述为它们预测地球表面每公里曲率的大小。“地平理论”会说每公里的曲率为0。现在我们都知道，这种理论是错误的。但从某种意义上说，它又很接近真理。正如阿西莫夫（1989）所说：

> 亚里士多德之后的一个世纪，古希腊的另一位哲学家埃拉托塞尼斯（Eratosthenes）指出，太阳在不同纬度上投射不同长度的影子（如果地球是平面的，所有的影子都应该一样长）。根据影子长度的不同，他计算出地球的周长为2.5万英里，那么这个球体曲率是0.000126度/英里。正如你所见，这个数值非常接近0……这从0到0.000126的差别解释了为何我们用了如此长的时间，才放弃“地球是平的”这一观念，并转而相信地球是球状的。提醒你一下，即使是像0～ 0.000126之间这样细小的差别也是至关重要的。失之毫厘，谬以千里。如果这点小差别没有被考虑到，如果地球被认为是一个平面，而不是一个球，那么我们将无法精确地绘制地球上大面积区域的地图（pp. 39-40）。

当然，科学并没有止步于“地球是球体”这一理论。我们早先讨论过，科学家们一直在尝试尽量改进他们的理论，并挑战当前知识的局限。例如，牛顿的引力理论预言地球并不是完美的球形，这个预言确实被证实了。现在已经证明，地球在赤道附近略微凸起，而在两极附近略微扁平。这是个被叫做“扁球体”的形状。地球从北极到南极的直径是7900英里，赤道直径是7927英里。所以，地球的曲率并不是一个常数（像一个理想的圆球那样），而是在每英里上有约7.973英寸到8.027英寸的微小变化。正如阿西莫夫（1989）所言：“从球体到扁球体的修正比从平面到圆的修正要小得多。因此，虽然‘地球是球状的’这一理解有误，但严格地说，它没有错到‘地球是平的’那种程度。”（p.41）

阿西莫夫关于地球形状的例子为我们展示了科学家们使用“错误”、“误差”和“证伪”这些术语的不同情境。这些术语并不是说被检验的理论错得一无是处，这些理论仅仅是不完善的。所以，当科学家强调说理论是暂时性的、可能被未来的研究发现所修正的时候，他们所指的就是例子当中的情形。当科学家相信地球是球状的时候，他们认

识到在未来某一天，这个理论需要在细节上进行修正。无论如何，从球体到扁球体的变化维持了地球是一个球体的“大体正确性”。我们绝不会在某天醒来突然发现它其实是一个立方体。临床心理学家斯科特·利连恩费德（Scott Lilienfeld, 2005）向心理学专业的学生这样介绍阿西莫夫的观点：

> 当向学生解释心理学知识本来就是暂时性的、可以被修正的时候，有些学生会错误地得出结论，认为真正的知识是不存在的。这种观点在某些后现代主义圈子里非常流行，它忽视了对不同确定程度的知识的区分。虽然绝对的肯定在科学中无法实现，但一些科学理论，如达尔文的自然选择学说，已经被极好地证实了，而其他一些理论，如支持占星术的理论，已经遭到了有力的驳斥。还有一些理论，如认知失调理论，仍处在科学争议当中。因此，科学理论是个确定程度不同的连续体：有些已经成为了确定的事实，另外一些则被完全地证伪了。对于科学问题，方法论上的怀疑主义并不产生完全确定的答案（原则上说，这些答案可能会被新的证据推翻），这个事实并不意味着知识是不存在的，只是说知识是暂时性的。（p.49）

“科学必须产生确定的知识”这一错误信条常常被用来攻击科学。古生物学者尼尔·舒宾（Neil Shubin）描述了神创论者如何利用这一策略。在和科学作家纳塔利·安吉尔（Natalie Angier, 2007）的访谈中，舒宾说：“神创论们首先试图将科学描绘成铁板一块的事实，然后攻击那些‘确定’中不是特别确定的事情。他们叫嚷着‘哈！你自己都理不清思路了吧，你太不不靠谱了。那我们为什么要相信你所说的一切呢？’其实是他们首先竖起了一个‘科学毫无瑕疵’的稻草人当靶子。”（p.20）

小结

科学家们提到“可解的问题”时，通常指的是“可检验的理论”。“可检验的理论”的定义在科学上是非常明确的：这个理论是有可能被证伪的。如果一个理论不可证伪，并且和自然界的真实事件没有关联，那么它就是无用的。心理学里一直充斥着不可证伪的理论，这也正是心理学发展缓慢的原因之一。

好的理论能够作出具体的预测，具有高度的可证伪性。相比于一个不精确的预测，如果一个明确具体的预测得到证实，就会为产生这个预测的理论提供更大的支持。简言之，可证伪性原则的一个含义就是，并非所有理论的验证都具有同样的价值。可证伪性越高，预测越具体，得到证实的理论就越受青睐。即使预测并没有得到证实（比如它们被证伪了），可证伪性对于理论的发展也是有用的。一个被证伪的预测说明，原有理论要么应当抛弃，要么需要进行改变，以解释不一致的数据。正是通过这种由被证伪的预测所引发的理论修正，像心理学这样的科学才能逐步向真理逼近。

Chapter 3
操作主义和本质主义："但是，博士，这到底是什么意思？"

物理学家真正理解地心引力是什么吗？我的意思是"真正"。他们知道"地心引力"这个术语的真正含义是什么？它的内在本质是什么？说到地心引力时，最终所要表达的意思是什么？说到底，它究竟是什么？类似这样的问题反映了一种科学观点，哲学家波普尔称其为"本质主义"。这种观点认为，从内在本质或者本质属性的角度对现象作出最终解释，才算得上是好的科学理论。支持这种观点的人通常也相信，无法对现象作出最终解释的任何理论都是无用的，这样的理论不能反映真实的内在情况，不能反映世界存在方式的本质。本章，我们将讨论为什么科学不去回答本质主义者的问题，而是通过对概念进行操作性定义得以进步。

为什么科学家不是本质主义者

事实上，科学家并不企图获得本质主义者所追求的那类知识。从这一意义上讲，对于本章一开始提出的问题的正确回答是：科学家不知道地心引力是什么。科学并不试图回答关于宇宙的“终极”问题，彼·得米德瓦（1984）曾写道：

> 确实存在那些科学不能回答并且在科学发展的可预见的范围之内也不可能得到答案的问题。比如，那些孩子们会提出的问题——“终极问题”……我能想到的这样的问题有：世界是如何开始的？我们来到这世间是为了什么？生活的意义是什么？（p. 66）

> 然而，即使科学不能回答终极问题，也不意味着必须接受其他的答案；也不能理所当然地认为，既然这类终极问题能被提出，就一定能够被回答。就我们目前的理解力而言，这类问题是无从回答的。（p. 60）

> 但是，最终就其能回答的那类问题而言，科学的潜力是无穷的……没有什么可以阻挡或终止科学的发展，除了诸如缺乏勇气之类的道德方面的缺陷。（p. 86）

科学家之所以质疑那些自称为终极问题给出绝对答案的人、理论或者观念体系，一个原因就是科学家认为终极问题是无法回答的。科学家并不会宣称他们可以提供完美的知识；科学的独特优势并不在于它是一个不会犯错的过程，而在于它提供了一种消除错误的方式，它能不断消除我们认识中的错误。再者说，自称完美或绝对知识的主张及做法，却往往会阻碍人们的探索。自由而开放地探索知识是科学活动的一个先决条件。科学家们总是在怀疑那些号称已经找到问题最终答案的言论。

本质主义者喜欢咬文嚼字

本质主义者的态度通常有一种表现：在探求知识之前，过于关注术语或概念的定义。“但是，我们必须首先界定我们的术语”是本质主

义者常用的一个口号。“某理论性概念的真正含义是什么”这种理念似乎意味着，当一个词被当作理论中的概念使用之前，我们必须对这个词的使用所涉及的所有潜在语言问题有一个全面且清晰的理解。事实上，这正好与科学家的工作方式相反。在对物理世界开展研究之前，物理学家不会花费气力讨论如何使用“能量”一词，或者当我们讨论物质的基本组成时，“粒子”一词是否真正表达了我们要表达的本质含义。

在科学领域里，确定某概念的意义，是在与该术语有关的现象得到一定程度的研究之后，而非研究之前。一个精确的概念性术语来自科学过程中固有的那种数据和理论间的相互作用，而不是关于语言用法的辩论。本质主义者让我们陷入无休止的文字争论，而许多科学家坚信这样的文字游戏使我们脱离了事物的实质。例如，对于“生命一词的真正含义是什么”这个问题，两个生物学家给出的令人惊讶的回答是“没有什么真正的含义，它只是足够好地满足我们生物学家工作需要的一种用法，并不是争论或辩驳的主题”（Medawar & Medawar, 1983, pp. 66-67）。总之，科学家的目的是解释现象，而非对措词进行分析。在所有的科学学科里，进步的关键在于放弃本质主义，接受操作主义。这正是本章中我们探讨的主题。

操作主义者将概念和可观测的事件联系在一起

那么，如果不是来自语言文字的争论，科学中概念的含义又来自哪里呢？正确使用某一科学概念的标准是什么？为了回答这些问题，我们必须讨论操作主义。它对于在科学领域中建构理论至关重要，尤其对于评估心理学中的理论及观念具有重要作用。

尽管操作主义形式多样，但是对于科学信息的使用者来说，用最宽广的思路去思考操作主义是最有效的。“操作主义”只是这样一种思想：科学理论里的概念必须立足于可观测事件，或与可观测事件相关联，而这些可观测事件是可以被测量的。将概念与可观测事件相联系的是概念的操作性定义，这使概念公开化了。操作性定义使得概念从个人化的感觉和直觉中分离出来，并且允许任何实施可测量操作的人对概念进行检验。

例如，把“饥饿”这个概念定义为“我胃里不好受的感觉”，并不是一个操作性定义，因为它与个人对于“不好受的感觉”的体验相联系，因此不能被其他观察者知悉。相反，涉及一些可测量的食物剥夺时间或者像血糖水平这样的生理指标的定义才是可操作性的，因为它包含

了任何人都可以实施的可观测的测量。同样，心理学家不同意将“焦虑”定义为“我不时会感到的不舒服和紧张”，而是必须用像问卷和生理指标测量这样的操作来定义概念。上述那个定义是个人对身体状况的解释，他人无法复制；而后者则是把概念放在公共科学领域进行解释。

在科学领域里，定义一个概念靠的是一系列操作，而非单独的行为事件或任务，意识到这点非常重要。相反，一些差别细微的任务和行为事件通常聚合在一个概念上（在第8章我们将会更多地讨论聚合性操作）。例如，教育心理学家根据利用诸如《伍德库克阅读能力量表》之类（Woodcock, 2011）包含一系列任务的标准化工具测得的成绩来定义“阅读能力”这个概念。该量表测出的阅读能力总分包含了一些不同分量表测得的能力指标。这些分量表测查的能力稍有不同，例如，阅读一篇文章、想出一个合适的单词在文章中填空、写出一个词的同义词、独立拼读一个较难的词，等等。所有这些任务上的表现综合地定义了“阅读能力”这个概念。操作性定义促使我们认真地、经验性地思考我们如何定义一个概念，所谓经验性，是指要根据我们对真实世界的观察。试想，我们要给一个看起来相当简单的概念“打字能力”下一个操作性定义。想象一下你这么做是为了比较两种打字教学方法的优劣。思考一下你所要做的所有决定。当然，你想要测量“打字速度”。但是要打多长的一段文章呢？仅有100个单词的文章可能太短，而10000个单词的文章又似乎太长。那么到底多长才算好呢？打字速度维持多久才最符合我们对打字能力这一概念的理论建构？用什么类型的文章来测试呢？文章是否要包含数字、公式和不常见的间距？我们如何处理错误？当我们测量打字能力的时候，时间和错误似乎都应被考虑在内，但是如果把这两个指标同时考虑进去的话，要如何来计算一个总分呢？我们想要让时间和错误具有相同的权重，还是一个比另一个更重要？寻求一个好的操作性定义会迫使你认真考虑所有这一切：它会让你对如何将“打字能力”进行概念化做一番透彻的思考。

考虑美国食品和药物管理局（the Food and Drug Administration, FDA）的任务，他们决定什么是各类食物“不可接受”的污染水平，什么被认为是食品中“不可避免的缺陷”（Levy, 2009）。作为FDA这样的联邦机构也不能够依靠主观作出判断。这需要在检查每种食物时，对每种污染物的操作定义进行严格规范。例如，它构建了以下类型的操作定义（Levy, 2009）：番茄汁中“不可接受”的污染水平是每100克中超过10个苍蝇卵，蘑菇中“不可接受”的污染水平是每100克有五个以上的2毫米以上的蛆。简单粗暴却极具操作性！

信度和效度

对一个概念下操作定义包括测量：通过一些法则给观测结果赋值。科学作家查尔斯·塞费（Charles Seife, 2010）指出，当我们在测量使用数字时，我们才突然开始重视它们。他认为，如果数字不作为抽象符号使用，非数学家几乎不关注数字的性质。我们不会关注数字5本身。但是一旦数字5变成“5磅”或“5美元”，或是“通货膨胀率为5%”，又或是“智商5分”——突然之间，我们便开始关注这了。塞费（2010）说：“没有单位的数字是抽象而缥缈的。有了单位，它获得了意义，但是同时它也失掉了自身的纯粹性”（p.9）。赛费所说“失去纯度”的意思是一旦涉及测量，就是说给数字附加单位，我们突然关心数字的“正确”性质。在科学上，什么是数字的“正确”性质呢？这个问题的答案是，在科学上，数字应该有的“正确”性质是信度和效度。

概念的操作性定义要想有用，必须同时具备信度和效度。信度是指测量工具的一致性。如果你对同一概念进行多次测评，是否能够得到相同的测量结果。信度的科学概念很容易理解，因为它与常识中的定义以及字典里的定义非常相似：“任何总能够产生相同结果的系统所具备的一种属性”。

试想一下，一个外行人士会如何评价一件事是否可信呢？想象一个每天早上要赶公共汽车从新泽西去曼哈顿上班的人。按照时间表，公共汽车每天应该在上午7:20到达此人等车的站点。在一个星期中，如果公共汽车到达的时间分别是7:20、7:21、7:20、7:19和7:20, 那么我们就可以说在那一周汽车的到达时间是可信的，如果下周汽车到达的时间分别是7:35、7:10、7:45、7:55和7:05, 那么我们就可以说在那一周汽车的到达时间是非常不可信的。

在科学领域中，一个操作性定义的信度以类似的方式来评估。如果我们多次测量同一概念得到的结果是近似的，那么我们就说测量工具表现出较高的信度。如果在同一星期的周一、周三和周五，用同一IQ测验的不同版本测量同一个人的智力，得到的分数分别是110、109、110, 那么我们可以说这一IQ测试是非常有信度的。相反，如果三个测试分数分别是89、130和105, 那么我们就可以说这一IQ测试没有显示出高信度。有一些专门的统计方法可以评估不同类型的测量工具的信度，所有标准的方法论入门教材中都有介绍。

但是请记住，信度仅仅是指前后一致，而不包括其他内容。对于一个操作性定义而言，仅有信度是不够的，信度是一个必要而非充分条件。作为一个好的操作性定义，操作必须被证明对于概念来说是有

效的测量。“结构效度”这个术语是指一个测量工具（操作性定义）是否测量了它本应测量的内容。保罗·考兹比教授（Cozby, 2012）在其所著的方法论教材中为我们讲述了一个只有信度而没有效度的搞笑例子。假设你想测测自己的智力，测试者让你站到一个类似鞋码器的测试仪器上，然后仪器给出一个读数。当然，你会认为这是一个笑话。但是请注意，这个测量工具可以显示许多类型的信度，而这些信度在方法论教材中都会讨论到。这个仪器在星期一、星期三和星期五会呈现出相当一致的读数（这称之为“重测信度”），并且无论谁操作它，它都会给出一样的读数（称之为“评分者信度”）。

用鞋码器来测量智力，其问题不在于信度（这是有信度的），而在于效度。它不是一个测量其本应测量的概念（智力）的合理方式。断定它不是测量智力的有效方式的证据之一，就是我们发现它和其他一些被认为与智力相关的变量无关。鞋码器的测量结果与学业成就无关，与脑功能的神经生理学测量无关，与职场成功无关，与认知心理学家提出的信息加工效率的指标无关；相反，真正的智力测验与所有这一切都有关（Deary, 2001; Deary et al., 2008; Duncan et al., 2008; Flynn, 2007; Geary, 2005; Hunt & Carlson, 2007; Lubinski, 2004）。在心理学领域，真正的智力测验要兼顾效度与信度，而智力的鞋码器测验只有信度而没有效度。

在这一点上，你可能想知道信度和效度的其他组合方式是否可行。因此，让我来重申一下我们的立场。在操作性定义中，我们寻求信度和效度兼备，因此高信度和高效度结合才是理想的目标。我们刚刚讨论了鞋码的IQ测试，目的是论证高信度和低效度是没用的。第三种情况是低信度和低效度，这绝对没有用，因此不值得讨论。但是你可能想知道第四种，也就是最后一种可能的组合方式：如果高效度和低信度又怎么样呢？答案是和低效度和高信度的例子（鞋码器例子）一样，这种组合也是没用的。事实上，更准确的说法是，这类情况压根儿不可能出现。因为，如果不能进行可信的测量，你根本无法宣称测量是有效的。

当形成有效的操作定义时，精确我们要测量的概念是十分重要的。例如，美国橄榄球联赛（NFL）中评价四分卫时会用到“传球评分”（passer rating）这个指标（Sielski, 2010）。意识到这个指标被精确地确定为“传球评分”很重要。也就是说，它很明确不是四分卫评分指数。那是因为“传球评分”的操作定义把只是四分卫传球而不是四分卫做的每件事情考虑了进去。具体地说，传球率是一个数学公式，包括四个方面：完成传球百分比、每次传球尝试的码数、每次传球尝试的触地得分、每次传球尝试的拦截。传球评分指数统计数值不包

括：四分卫跑的码数、呼叫战术能力、胜负记录、在争球线后抱摔、掉球和其他一些可量化的四分卫变量。为此，在另一个操作定义下的叫做“四分卫总评分”的概念产生了。

直接和间接的操作性定义

概念和可观测的操作之间的联系，在直接和间接性程度上变化很大。很少有科学概念几乎完全是通过可观测的操作来定义的。大部分概念的定义采用更为间接的方式。例如，一些概念的使用既取决于一系列的操作，又取决于它和其他概念之间的特殊关系。最后，还有一些概念不通过可观测的操作直接定义，而是通过它与另外一些概念间的关系来定义的。这种概念有时被称为“潜在概念”，在心理学中非常普遍。

举个例子来说，许多研究关注所谓的A型行为模式，因为它与冠心病的发病率有关（Chida & Hamer, 2008; Martin et al., 2011; Matthews, 2005; Smith, 2003; Suls Bunde, 2005）。在第8章中，我们将会更加详细地讨论A型行为模式。但是，这里重点要说的是，A型行为模式实际是通过一系列二级概念来定义的：强烈的竞争欲望、潜在的敌意、赶时间行为、达成目标的强烈驱力，等等。然而，每一个用于界定A型行为模式特征的概念本身也都需要操作性定义。事实上，研究者们已经为对每个概念进行操作性定义而付出了很多努力。我们讨论的要点是，A型行为模式是一个复杂的概念，它并不是被操作所直接定义的。相反，该概念与其他一些各自具有操作性定义的概念联系在一起。

A型行为模式提供了一个间接操作性定义的例子。在临床心理学上，有一个相似的概念定义是痛苦耐受性（Zvolensky, Vujanovic, Bernstein, & Leyro, 2010）。这个整体的概念由几个简单的依赖于操作测量的子概念构成：对不确定性的容忍度、歧义容忍度、挫败容忍度、负面情绪容忍度以及身体不适容忍度。

简而言之，尽管不同的概念与可观测操作的联系程度各有不同，但所有的概念都在一定程度上通过其与可观测操作之间的联系来获得意义。

科学概念的演进

一个科学概念的定义并不是固定不变的，而是随着相关观测结果的不断丰富而发生变化。意识到这一点非常重要。如果一个概念的原始操作性定义在理论上被证明是无效的，那么该定义就会被抛弃，以另外一套定义的操作取而代之。这样，随着相关知识的积累，科学概念不断演进，其抽象性逐渐增加。例如，在一段时间里，人们认为电子是一个围绕原子核旋转的带负电的微小球体。而如今，电子被视为在特定实验条件下，具有似波特性的概率密度函数。

在心理学领域，智力概念的发展提供了一个类似的例子。起先，智力仅有一个严格的操作性定义：智力是通过心理功能测验所测到的东西。随着实验证据的不断积累，智力被证明与学业成就、学习、脑损伤、神经生理学及其他行为和生物学变量有关，这一概念在逐渐丰富的同时又得到了提炼（Deary, Penke, & Johnson, 2010; Duncan et al., 2008; Sternberg & Kaufman, 2011）。现在看来，在定义智力概念时，最好用一种高等级的建构，通过多种更为具体的信息加工操作来定义。当然，这些假设的信息加工过程应该具备更为直接的操作性定义，可以用可测量的指标来表述。

人类记忆理论中的概念也以同样的方式发展。现代心理学家很少使用类似“记忆”或“遗忘”这样的笼统概念；相反，他们测量那些可以进行明确定义的记忆子过程，如短时听觉记忆、符号存储、语义记忆以及情景记忆。传统的“记忆”或“遗忘”的概念通过更加明确的操作性概念得到了细化。

因此，理论术语的用法在科学实践中不断演进，而不是在针对文字意义的争论中获得发展。这是科学的操作态度和本质主义者在追求绝对定义之间最显著的区别。哲学家保罗·邱吉兰德（Churchland, 1988）强调，在科学中，对概念的定义不是源于文字界定，而是源于与之相关的观察和其他概念：

> 要想完全理解“电场”这个概念，我们就必须熟悉这一表述所处的理论原则体系，它们会共同告诉我们，电场是什么、做什么。这是一个典型的例子。通常来讲，理论性术语的意义不是从单一的、具体描述其所适用的必要充分条件的定义中获得的，它们往往通过所在的理论原则体系而被间接地定义。（p. 56）

随着科学概念的演进，概念常常与许多不同的理论体系交织在一起，并且获得多种操作性定义。这种情况的出现并不是因为概念本身出了问题。例如，许多人认为心理学不可信，因为心理学中许多重要的理论概念——例如智力，可以用不止一种方法来操作化和概念化。但这种情形并非心理学所独有，也不是一件令人绝望或束手无策的事

情。事实上，在科学领域里，这种情况是普遍存在的。例如，“热”既可以从热力学理论，也可以从动力学理论的角度来概念化。物理学并未因此遭到贬斥。想想电子，它的许多特性都是以波的概念来解释的。可是，如果将其视为粒子，它的另外一些属性则更好理解。到目前为止，还没有一个人会因为物理学中存在着这些多重概念化现象就提出要抛弃它。

2006年，人们在这一点上有了深刻的认识。当时媒体报道国际天文联合会（International Astronomical Union）对“行星”的定义重新进行了操作化，并将冥王星排除在外（Adler, 2006; Brown, 2010）。像“行星”这样基础的概念都会发生变化，确实让公众大为震惊，但这在科学领域是司空见惯的。在行星定义这件事儿上，一些天文学者倾向于强调天文主体的构成和地质因素；另一些天文学者则强调行星的动力属性，例如它们的轨道和重力影响。在前一组天文学家的操作性定义中，冥王星是行星，而在后一组天文学家的操作性定义中，冥王星则不是行星。不同的操作性定义反映的不是天文学界的混乱，它们只是反映从不同角度定义概念的方式。在心理学上也是一样，一个概念也有许多可选择的操作性定义。有些事物难以定义，并不代表没有可研究的东西。

心理学领域的操作性定义

许多人在思考物理学或化学的时候，能够理解操作主义的必要性。他们知道，如果科学家准备谈论某一类型的化学反应、能量或者磁场，就必须有相应的方法来测量。不幸的是，当人们谈到心理学的时候，却经常无法认识到操作主义的必要性。为什么人们没有同样地认识到这一显而易见的事实：为了成为科学理论中有用的解释体系，心理学术语必须被直接或间接地操作化定义？

人们对心理学产生误解的原因之一，就是心理学上所说的“预设偏见”。在第1章中，我们提到过这个问题。人们不会因为执着于某种关于岩石性质的信念来研究地质学，而在心理学中，情况就大为不同了。我们每个人都有关于人格和人类行为的直觉理论，我们用它们来“解释”自己以及他人的行为。我们所有的个人心理学理论里都包含着理论性概念（例如聪明、攻击和焦虑）。因此人们会很自然地发问：“为何我们必须接受一些其他的定义？“尽管这种态度从表面上看来是合理的，但对于任何致力于理解人类行为的科学来说，它都是一个巨大的障碍，也是公众对心理学产生困惑的一个原因。

误解产生的最主要原因，就是心理学中的许多专业概念都是用日常用语来表达的，这也是媒体在准确呈现心理学成果方面最大的障碍。这些日常用语为大量误解的传播敞开了大门。外行人很少意识到，当心理学家把“智力”、“焦虑”、“攻击”、“依恋”等词语当作理论性概念来使用时，它们的含义和大众平常所说的意思不一定相同。

从之前关于操作主义的讨论中，就能看出这种区别的本质。当在心理学理论中使用如“智力”、“焦虑”这些术语时，它们直接或间接的操作性定义决定了它们的正确用法。那些定义常常具有高度技术性，通常具有特定意义，并且在许多方面都不同于这些术语在日常生活中的运用。例如，当我们听到“对大样本的认知任务进行因素分析所得到的第一个主成分”这段话时，许多人都意识不到它是术语“智力”的操作性定义。

同样，如果外行人使用术语“抑郁”，那就意味着“感觉糟透了”。相反，在《精神疾病诊断与统计手册》（Diagnostic and Statistical

Manual of Mental Disorders）中，对抑郁症的专门定义占用了超过12页的篇幅（American Psychiatric Association, 2000），并且与“感觉糟透了”有着很大的区别。临床心理学家所谓的抑郁，并不等同于外行人所说的抑郁（Hollon, Thase, & Markowitz, 2002）。在其他科学领域也都存在同样的问题，尽管可能没有心理学这么严重。回想一下前面对“生命”概念的讨论。正如米德瓦等（1983）指出的，问题在于，像科学中的其他专业术语一样，“生命”一词出自人们的日常用语，但在科学场合中的含义已远不同于日常谈话中的用法（p. 66）。

物理学家丽萨·兰道（Randall, 2005）曾讨论过这类问题如何阻碍了公众对物理学的理解。她指出爱因斯坦相对论中的“相对性”一词被公众理解为“绝对是不存在的，因为任何事物都是相对的”，而事实上，该理论的意思正好相反！兰道指出，其实爱因斯坦的相对论都是有关事物的不变和绝对属性的。实际上，爱因斯坦也曾坦陈这一理论如果被命名为“恒定论”可能会比“相对论”更加贴切，但是“相对性”一词的地位在当时很快就已深入人心了。兰道继续指出，即使在物理学中，模糊的词语选择也是造成某些误解的根源，科学家经常使用一些口语化的术语，但外行人无法区分这些词语的用法是通俗意义上的还是技术意义上的。在心理学中也是如此。当心理学家和外行人用同一词语来表达不同含义的时候，他们之间常常产生误解。如果有新的词语产生出来用以描述心理概念，这样的困扰可能会少一些。有时会有这样的词语出现，正如物理学家有了“尔格”和“焦耳”一样，心理学家有了“失调”和“编码”，这些词不是凭空编造的，但在日常用语中比较生僻，从而可以防止混淆。

“但是，”外行人可能提出这样的反对，“为什么心理学家这样折磨我们？为什么有这么多新的术语、高度专业性的定义、生僻的词语？为什么我们需要这些？为什么我对‘智力’概念的定义得不到他们的认可呢？”

在这里，我们来看一个对心理学研究有严重误解的例子——这一误解经常出现在关于心理学研究的媒体报道中。一份全国性的报纸以“你能用一般人听得懂的话重新说一遍吗？”为标题报道了1996年美国心理学会的一次会议，并说“心理学家所用的语言只有他们自己能听懂”。该文嘲讽了在会上报告的一份题为《用Gf-Gc理论解释对WJ-R和KAIT的联合因素分析》的论文。尽管记者表示他“不敢贸然猜测这个标题的真正意思”，但几乎所有接受过培训的心理学家都能理解这个标题是有关智力测验理论方面新进展的。的确如此。Gf-Gc理论是智力理论方面的一个技术性进展，记者基本不会听说过这个概念——就如同我们不会期望这位记者知道物理学家最近刚发现的一种基本粒

子的细节一样。可是有时候，记者对科学术语的无知（这是完全可以理解的）却对现代心理学造成了负面影响。当话题是物理学时，记者似乎知道是他们的无知妨碍了他们的理解。但是，当话题是心理学时，他们就表现得好像心理学家要为他们的不懂负责一样。

我们来看看问题的症结所在。解决它的第一步，就是强调我们已经讨论过的一个观点：操作主义不是心理学所独有的，它是所有科学门类的特征。大多数情况下，我们很容易接受它，理解它的显而易见的本质。如果一位科学家是研究放射性的，我们会理所当然地认为他肯定有办法测量此种现象——其他研究者也能使用该方法获得相同的结果。操作定义因此导致科学的公开化，而公开化是科学的关键特征之一。如果两个科学家对同一个操作性定义达成一致，其中一个人就可以用它去复制另一个人的结果。但是，在其他情况下看来显而易见的事情，在我们谈到心理学的时候却不怎么明晰了。人们经常意识不到“智力”和“焦虑”这些概念的操作性定义的必要性，因为我们总是在使用这些术语，难道我们还不“知道”它们是什么意思吗？

答案是：是的，我们确实不知道它们是什么意思——不是从科学家必须知道的意义上，而是从公众的意义上。一个科学家必须通过如下方式“知道”智力是什么意思：他必须精确地定义一种方法，使其他实验者能够以完全相同的方法测量这一概念，并且得到有关此概念的相同结论。就其明确性和精确性来说，这与日常交谈中为了实现相互理解而使用的模糊语言间有很大的差别。

作为人性化力量的操作主义

如果过分依赖于我们“知道”的东西，肯定会产生问题，这个问题同样困扰着所有的直觉（非经验主义的）信仰体系。关于某个事物，你所“知道”的和张三、李四所知道的可能并不一样，我们如何决定谁是正确的呢？你或许会说：“我强烈地感觉到我所知道的是正确的。”但是，如果张三的观点和你有出入，但比你拥有更强烈的感受呢？李四的观点与你俩都不同，也宣称自己是正确的，因为他的感受甚至比张三还要强烈。

讲这个简单的小段子，仅仅是想阐述科学知识的一个基本特点，它在人类历史中是一股重要的人性化力量：在科学中，知识的正确与否并不取决于个体提出主张时自己的肯定程度。所有建立在“直觉”基础上的信念体系都有一个共同的问题，即当出现矛盾观点时，它们缺乏一种机制来判别哪个是对的，哪个是错的。因为人人都凭直觉认为

自己是对的，但当大家的直觉观点发生冲突时，我们该如何决定谁正确呢？令人悲哀的是，历史表明，这种冲突的结果通常是权力斗争。

一些人错误地宣称，心理学的操作取向使人们丧失了人性，而且我们应该把我们关于人类的观点建立在直觉基础之上。心理学家唐纳德·布罗德本特（Donald Broadbent）在1973年论证说，真正人性化的观点是将关于人类的理论观点建立在可观测的行为基础上，而不是以理论者的直觉为基础：

> 除非我亲眼看到别人在特定情况下亲自做了或说了什么，否则无法对别人作出判断……实证主义的方法是一种调和差异的方式。如果拒绝这一方式，那么处理争论的唯一方式就是面红耳赤地争辩了（p. 206）。

因此，科学中人性化的力量就是让知识公开化，让任何有冲突的观点都能以一种双方都接受的方式得以检验。回想第1章中提到的“重复”的概念。这让我们可以通过一种大家都事先同意的、平和的方式来从理论中进行选择。科学的公共性本质在很大程度上依赖于操作主义的理念。通过对概念操作化的界定，概念进入了公共的领域——任何人都可以对其进行批判、验证、改进或否定。

心理学概念不能以某些人的个人定义为基础，因为这类定义可能是不常见的、个人化的或者模糊的。由于这个原因，心理学必须摒弃所有对概念所做出的个人化定义（就像物理学拒绝对能量的个人化定义，气象学拒绝云的个人化定义），而坚持公众可以知悉的定义，这种定义用操作来界定概念，并确保了任何一个接受过适当训练并拥有适当设备的人都可以实施这些操作。就摒弃个人化定义而言，心理学并没有将外行人拒之门外，而是将这一领域向公众敞开，就像所有学科那样，以期寻求所有人都可以共享的、普遍的、公众可以利用的知识。

只有当概念以操作性定义为基础，并且不关注于本质主义者所讨论的文字意义时，这类具有公众可用性的知识才能够用来解决人类的问题。例如，蒙克（Monk, 1990）描述了第二次世界大战期间“创伤性休克”这个概念是如何在医学领域产生问题的。一些医师对此症状的诊断依据是过高的血红细胞浓度，并认为其原因在于血液中的血浆渗透到了组织中。其他医师诊断“创伤性休克”则根据低血压、皮肤苍白和脉搏过速。换言之，医生们对这一概念的操作性定义是不一致的（甚至是带有个人色彩的）。因此，英国医学研究会的格兰特（Grant）医生建议说，“创伤性休克”这个概念应该被抛弃，并且对伤者进行详细观察时也不使用这个术语……由于在诊断方面缺乏共同

的基础，无法对各种治疗措施的效果进行评估（Monk, 1990, pp. 445-446）。换句话说，这种概念弊大于利，因为缺乏一个获得普遍认同的定义使之成为公共知识（也就是被广泛地分享与认同）。

有时候，在科学领域中，概念意义的改变会导致对这一概念的科学理解与外行人士的理解产生冲突。法伯和邱吉兰德（Farber & Churchland, 1995）讨论过一个关于“火”这一概念的例子。经典的概念是这样定义火的：“不仅是含碳物质的燃烧，而且还包含了太阳及各种星体上的活动（实际是核聚变）、闪电（实际上是电引起的白热化现象）、北极光（实际是光谱发射）和萤火虫的闪光（实际上是发出磷光）。在现代概念体系中，这些现象都与氧化无关，因此和木材燃烧不属于同一类型。另外，有一些现象原本认为是与燃烧没有任何关系（由于那时放热被认为是燃烧的本质特征）的过程——如生锈、锈蚀和新陈代谢——却被证实属于氧化现象。”（p. 1296）总之，氧化的原则使得篝火和生锈联系了起来，而将闪电与它们区分开来。对于科学家而言，这也许是一个进步的标志，但却让外行人士感到迷惑和无所适从了。

本质主义者问题和对心理学的误解

许多人在接触心理学时，放弃操作主义观点的另一个原因是，他们想为特定的人类问题找出本质主义的答案。回想本章开头提出的问题：“地心引力”这个术语的真正含义是什么？它的内在本质是什么？在谈到地心引力一词时，我们到底指的是什么呢？大部分人认为这些问题需要绝对性的知识，需要理解现象的潜在本质，而物理学当前的理论不能对这类问题提供答案。对关于物理科学近几百年来发展的通俗读物比较熟悉的人都能意识到，地心引力是一个高度复杂的理论建构，并且其概念性和操作性关系也处在不断变化之中。

可是，如果将上述问题中的“地心引力”全都换成“智力”，奇迹就出现了。那些问题此时立刻被赋予了重大意义。它们看起来是那么自然和富有深意，它们就是在寻求一个终极答案。可是当心理学家给出和物理学家一样的答案，即“智力是一个复杂的概念，它的意义是由测量它的操作以及它与其他概念之间的理论关系来界定的”时，却会被鄙视和指责为回避真实问题。

心理学所面临的一个难题就是，公众要求心理学去回答本质主义问题，而通常其他科学家并不需要回答类似的问题。这类要求常常导致人们贬低心理学领域已经取得的进步。尽管这类要求不能阻止这一

领域自身的发展——因为就像其他科学家一样，心理学家无视本质主义问题并继续他们的工作——但那些问题成了公众理解心理学的障碍。当一个不了解情况的批评家声称心理学没有取得进步时，公众就会迷惑了。这类责难极少遇到挑战，这也反映了本书序言中所述的不幸事实：对于心理学领域所取得的科学成就及其意义，公众的了解是极度匮乏的。当我们仔细审视那些对心理学的批评，不难发现它们通常归结于一点：心理学至今没有为它提出的问题提供终极答案。对于这种指责，心理学毫不犹豫地低头认罪——像所有其他科学学科一样。

一些人可能会很不舒服地发现，包括心理学在内，没有任何科学可以对本质主义问题作出回答。霍尔顿和罗勒（Holton & Roller, 1958）讨论过当外行人被告知物理学不能够回答本质主义问题时所表现出的那种不安。他们谈论的是与放射性衰变有关的现象：发生衰变的放射性元素的原子数量与时间是呈指数函数关系的。可是，这种函数并不能解释为什么放射性衰变现象会发生。它不能回答外行人为什么它会遵循这个函数，也不能回答放射性衰变到底是什么。霍尔顿和罗勒告诉我们："我们必须平静地接受现代科学的局限性，它并没有声称可以发现'事物究竟是什么'"（pp. 219-220）。科学作家罗伯特·怀特（Wright, 1988）解释说：

> 伊萨克·牛顿的地心引力理论有些地方不如尽人意……毕竟，"在一定距离外作用"如何实现？牛顿回避了这样的问题……自牛顿开始，物理学家们一直在效仿他的做法……物理学家们不再尝试解释为什么事物必须遵守电磁学规律或地心引力规律（p. 61）。

同样，那些为人类本性问题寻求本质主义答案的人若是诉诸心理学，注定将会失望。心理学不是宗教，它是一个试图对所有行为作出科学解释的广阔领域。因此，心理学现在的解释是对行为的暂时性的理论建构，就目前来说，这些建构在解释行为方面优于其他解释。这些建构在将来注定会被更好的、更接近事实的理论概念体系所取代。

在评估一个心理学理论的可证伪性时，操作性定义的理念是一个非常有用的工具。概念有没有直接或间接地建立在可观测操作的基础上，是识别不可证伪的理论的重要线索。所以，那些不严格的概念——理论学家不能为它们提供直接或间接的操作性联系——都应该引起怀疑。与之相关的是科学家称之为"节省"的原则。节省原则是指，当两个理论有同样的解释能力的时候，较为简单的理论（涉及更少的概念和概念性关联）胜出。原因是，拥有较少概念性关联的理论在将来的检验中会更具可证伪性。

深刻理解操作主义的原则，也有助于我们识别不具备科学意义的问题。例如，在我的电脑文件夹里有一篇来自于国际联合出版社的在线服务文章，标题为《动物会思考吗》。这篇文章讲述了动物行为方面最新的实验。文章中所引述的研究没有任何错误，但是很显然，这个标题仅仅是一个玩笑。这个标题的问题在于没有科学意义，没有关于“思考”的操作性标准。许多报纸的标题中存在类似的问题，比如“计算机会思考吗”没有操作性标准的话，这个问题也没有科学意义，但在鸡尾酒会上倒是可以派上大用场。

小结

操作性定义是利用可测量、可观察的操作来表述的概念定义。我们确信某个理论具有可证伪性的主要途径之一，就是确定理论中的关键概念具备可以用可重复性很强的行为观察来表述的操作性定义。操作性定义是让科学知识变得公开可检验的主要机制。这样的定义被置于公共领域，使其所界定的理论性概念能够接受所有人的检验，而不是像"直觉的"、非经验性的定义那样，只属于特定个体，检验它的机会并不向所有人开放。

由于心理学使用一些来源于日常生活的词语，如智力和焦虑，许多人对于这些术语的含义有着预设的想法，因此往往意识不到对这些术语进行操作性定义的必要性。心理学和所有其他科学门类一样，也需要对其术语进行操作性定义。可是，人们常常要求心理学家回答本质主义的问题（有关概念的纯粹深层本质的问题），而其他科学家就不必回答这类问题。没有科学能够回答这样的终极问题。心理学和其他科学门类一样，正在试图不断地完善其操作性定义，使理论概念能够更加准确地反映真实世界的原貌。

Chapter 4
见证和个案研究证据：安慰剂效应和了不起的兰迪

画面切换到《奥普拉脱口秀》——过去十年中最著名的电视脱口秀节目之一的脱口秀现场。今天的嘉宾是俄狄浦斯人类潜能研究所的所长阿尔弗雷德·庞蒂菲科特（Alfred Pontificate）博士。这位博士新提出了一个有关出生次序的激进理论，这一理论的基本理念是：个体的生命进程是被家庭互动所设定的，而家庭互动是由出生次序决定的。奥普拉鼓励观众对此理论进行提问。讨论不可避免地由最初的理论关注，转向了为观众个人生活中的重要事件作出解释。这位博士欣然应允。

例如，“博士，我的哥哥是个不要命的工作狂。他对妻子和家庭完全不管不顾，并且把与工作有关的问题看得比什么都重。他有溃疡和酗酒问题，但他拒不承认。他们家在近两年内从没过过一个真正意义上的假期。他的婚姻也快玩完了，但他似乎并不是特别在乎。他为什么要选择这样一种自我毁灭式的生活呢？”

博士反问道：“亲爱的，他在家中排行第几？”

“哦，他是子女中的老大。”

“这就对了，”博士说道，“这在生活中比较常见。我们在临床上经常见到这种现象。这类现象出现的深层次原因是，父母将自身的愿望和挫折都转移到他们第一个出生的孩子身上。通过愿望的这种无意识的转化过程，即使父母从未明确要求过孩子，孩子也在内化这些愿望和挫折。然后，通过这种我称之为‘期望上旋’的动力过程，父母的抱

负转化为孩子对于成功的病态渴求。”

当嘉宾挑战观众的信念时，《奥普拉脱口秀》的观众有时会提一些尖锐的问题，但当行为“专家”似乎是在印证观众的传统观念的时候，这种情况就很少发生。然而曾经有过那么一次，节目因为一位观众质疑嘉宾的主张而显得异彩纷呈。有一位热切而直率的观众当时正身处演播室，“但是请等一下，博士，”提问者开始了他的问题，“我哥哥也是家里的老大。我的父母把那个笨蛋送到哈佛，而让我去了一个将来能够成为一名牙医的两年制专科学校。但他们的‘神童’在一年之后就辍学了，跑到了科罗拉多州的山顶上。我们最后一次见到他时，他正在编篮子！我搞不懂你关于‘长子’的说法。”

这位观众使现场气氛骤然紧张，但是博士总是能够逢凶化吉：“哦，是的，我也曾经见过很多你哥哥这样的个案。是的，我经常可以在我的从业经历中遇到这样的人。他们的‘期望上旋’的动力过程发生中断，生成潜意识的要求来抵抗父母转化到他们身上的期望。这样的话，个体的生活规划会朝着与传统成就标准相反的方向发展。”一阵肃然的沉默之后，讨论转向了下一个“案例”。

这些场面我们都再熟悉不过了，只不过又是一个第2章讨论过的本杰明·拉什问题的例子罢了。关于出生次序的“理论”是在没有一个事例能够证明其“不成立”的思维框架下被构想出来的。由于它是一个不能证伪的理论，搬出再多能证明它的证据也没有意义，因为这个理论不能排除任何可能的情况。

然而，我们在本章所关注的并非这一理论本身，而在于那些用于支持它的证据。当被迫出示证据时，庞蒂菲科特博士搬出了他的“临床经验”或“个案研究”。这在媒体心理学领域是一个惯用的套路。脱口秀节目和通俗心理学图书中充斥着基于作者临床经验的心理学理论。他们通过这类渠道提供给公众的许多疗法，能够支持这些疗法的无非是那些曾接受治疗并认为得到了改善或被治愈了的人的个人见证。在本章中，我们将为心理学信息的消费者建立一个非常有用的原则：个案研究和见证作为评估心理学理论和治疗的证据是毫无价值的。

在本章中，我们将要证明这个原则为什么是正确的，并且还要讨论个案研究在心理学中的正确作用。

个案研究的地位

个案研究是对一个个体或很小的群体集中而详细的调查。个案研究的作用，很大程度上取决于科学探索在某个特定领域进展到什么程度。从个案研究或临床经验中获得的灵感，在特定问题的早期研究阶段或许比较有用，因为它们可以提示哪些变量需要进一步研究。个案研究在开启心理学新的研究领域方面起到过关键作用（Martin & Hull, 2006）。让·皮亚杰的工作就是很著名的例子。皮亚杰的研究提出了一种可能性，即儿童的思维并不只是成人思维的简易版或低级版，而是有其自身结构的。皮亚杰关于儿童思维的部分推测已经被证实，但很多还有待证实（Bjorklund, 2011; Goswami, 2008）。然而，对于我们这里的讨论来说，更重要的不是皮亚杰的哪些思想被证实了，而是要理解皮亚杰的个案研究尽管没有证实任何事情，但它为发展心理学家的研究提供了难以置信的广阔领域。第5章和第6章中所要介绍的相关研究和实验研究，为皮亚杰个案研究中提出的假设提供了或支持或否定的证据。

然而，当我们从科学研究的早期阶段（在此阶段个案研究可能是极为有用的）步入更为成熟的理论检验阶段之后，情况就大大不同了。由于个案研究在特定理论的检验中不能作为证实或证伪的证据，所以它在科学研究的后期不再有效。其原因就是：个案研究和见证叙述都是所谓的“孤立事件”，缺乏比较性信息，而这种信息对于排除其他可能的解释来说是必要的。

弗洛伊德工作的一个缺陷就是，他从不进入第二步，即从个案研究中建立的有趣假设转向对这些假设的真正检验（Dufresne, 2007）。研究弗洛伊德的重要作家之一弗兰克·萨洛韦（Frank Sulloway）说：“科学是一个两步走的过程。第一步是提出假设。弗洛伊德提出了一系列令人瞩目的、当时看上去非常有道理的假设，但他从未以严格的、科学所要求的方式采取关键的第二步。”（Dufresne, 2007, p.53）

见证叙述与个案研究相似，因为它们都是孤立事件。依赖见证叙述的问题在于，如果累积起来的见证几乎能够为每一种疗法提供支持，那么它也就不可能用来支持任何一种特定的疗法，因为所有相互

对立的疗法都有各自的见证。当然，我们想知道的是哪种疗法是最好的，但我们不能依据见证来决定。正如心理学家雷·尼克尔森（Ray Nickerson, 1998）在评论我们用以欺骗自己的认知过程时所说的那样，“江湖郎中的骗术往往得逞，是因为他们总能找到一些病人愿意为他们做见证，这些病人总是发自内心地告诉别人，他们自己的确从治疗中获益匪浅。”（p.192）例如，有大量的见证声称潜意识自助式录音带（用一种低于听觉阈限的信号制作出来的录音带）可以提高人的记忆力，甚至提高人的自尊，然而，在严格控制条件下进行的研究显示，这类录音带对记忆力或自尊没有任何改进（Lilienfeld et al, 2010）。

“其他可能的解释”这一理念，对于理解理论检验来说至关重要。实验设计的目标就是构建某一事件或现象，使其只能用某一种特定的理论来解释，而其他理论则解释不通。正如第2章在可证伪性上所说的，只有当我们收集的数据排除了其他可能的解释时，科学才能进步。科学为理论观点的自然选择创设了条件。

有些理论观点经过实证检验存活了下来，而另一些则被淘汰出局，凡保留下来的都更接近真理。但是，这是个慢工出细活的过程，各种理论观点都必须经过细致审查，以便发现哪些更接近真理。但是这一过程必须有所取舍：为支持某一特定理论所收集的数据，不能同时支持许多其他可能的解释。基于这一理由，科学家在他们的实验中设有控制组，或称为对照组，以期得到比较性信息。这样做的目的，就是为了能够在比较控制组与实验组的结果时，排除其他可能的解释。至于实验设计如何能做到这一点，将是后面几章的主题。

个案研究和见证叙述作为孤立的现象而存在，它们缺少必要的比较性信息，不能证明某一特定的理论或疗法更优越。因此，引用某个见证叙述或个案研究的结果来支持某一特定理论或疗法是错误的。如果这么做的那些人不指明他们所提供的所谓证据其实也适用于大量其他可能的解释，那他们就是在误导公众。简言之，针对某个现象的孤证具有高度的误导性。安慰剂效应的例子将更清晰具体地阐释这一论点。

为什么见证叙述毫无价值：安慰剂效应

几乎每种产生于医学和心理学的疗法都有一定数量的支持者，并且总能催生出一些发自内心认可其疗效的人。医学文献记载了猪牙齿、鳄鱼粪便、埃及木乃伊的粉末，以及很多更富想象力的东西都曾经具有疗效（Begley, 2008; Harrington, 2008）。事实上，人们早已熟知，仅仅暗示正在接受某种治疗，就足以使许多人感觉病情好转。

无论治疗是否有效，人们都会报告某种疗法曾对他们有所帮助，这种倾向被称为安慰剂效应（Begley, 2008; Benedetti, Carlino, & Pollo, 2011; Buhle, Stevens, Friedman, & Wager, 2012; Harrington, 2008; Novella, 2010）。安慰剂效应的概念在电影《绿野仙踪》中有绝佳的阐述。仙女并没有真的给铁皮人一个心脏，没有给稻草人一个大脑，也没有给狮子以勇气，但是他们都感觉更好了。实际上，直到近一百多年，医学才发展出较多具有确凿疗效证据的治疗方法，因此有人曾经这样说："本世纪以前，整个医学史只能说是安慰剂效应的历史罢了。"（Postman, 1988, p. 96）

我们可以通过对生物医学研究的考察来说明安慰剂效应这一概念。在生物医学研究中，所有的新药研究程序都必须包括对安慰剂效应的控制。一般来说，如果在一组病人身上试验一种新药，就要组建一个患同样病症的对等组，给他们服用等量不含任何药物的药剂（安慰剂）。两组病人都不知道他们吃的是什么药。这样，两组进行比较时，安慰剂效应——即给予病人任何一种新的治疗都会使他们感觉好些的这种倾向——就能得到控制了。仅仅报告百分之几的病人吃了新药后症状得以缓解是不够的，因为如果没有控制组的数据，就不知道报告症状缓解的病人是由于安慰剂效应，还是药物本身的疗效。

安慰剂效应在抑郁症治疗中是29%（即29%的病人服用安慰剂后报告症状缓解了），在十二指肠溃疡中是36%，在偏头痛中是29%，食道炎是27%（Cho, Hotopf, & Wessely, 2005）。将安慰剂效应与当下流行的抗抑郁剂百忧解结合起来，将会最大限度地发挥药物自身的作用（Kirsch et al., 2008）。安慰剂效应的效力是很强大的，以至于曾有报告说有人对安慰剂成瘾（Ernst & Abbot, 1999），这些人需要服用剂量越来越大的安慰剂来保持他们的健康状态！一个怪异的研究

（见Begley, 2008）发现，一组接受了虚假外科手术（有创口，但是没有进行实际手术）的被试，和真正接受手术的被试一样，报告了几乎相同程度的骨关节炎症状的缓解。与此类似，另一项研究发现许多进行肩袖肌腱撕裂手术的患者报告他们的疼痛消失了，虽然MRI显示结果表明他们的撕裂没有痊愈（Kolata, 2009）。

这些例子解释了为什么接近一半的医生报告说他们故意给患者开安慰剂（Tilburt, Emanuel, Kaptchuk, Curlin, & Miller, 2008）。最后，安慰剂效应会受到情境预期的调节。研究者证明（Waber, Shiv, Carmon, & Ariely, 2008），价格较贵的安慰剂比价格便宜的安慰剂更能缓解痛苦！

当然，在有关药物治疗的实际研究中，安慰剂控制并不只是一个什么都不含的药片，而是含有当前认为最有效的药用成分。实验比较的目的在于揭示，新药是不是比当前最有效的药还要好。

你每次吃处方药时都会得到安慰剂效应的提示信息，下次吃处方药的时候（如果你非常健康，就看看你祖母的药吧！），仔细查看一下药物附带的说明书（或者登录药品制造商的网站浏览一下），你将在药物问题说明里看到安慰剂效应的信息。例如，我吃一种叫作Imitrex（琥珀盐酸）的药物来缓解偏头痛。它附带的说明书告诉我：控制研究已经证实，在服用一定量的药物之后，57%的病人在两个小时内其症状得到了缓解（我就是这幸运的57%之一）。但是说明书同时告诉我，同样的研究显示，这类偏头痛中安慰剂效应是21%——有21%的人在服药后两小时内症状得到缓解，即使他们服用的药物里是中性材料而非琥珀盐酸。

安慰剂效应在心理治疗中也很常见（Lilienfeld, 2007）。许多有轻度和中度心理问题的人，在接受心理治疗后说他们的情况有所好转。然而控制研究证明：这一康复比例中，有相当一部分是由于安慰剂效应和时间推移这两个因素共同作用的结果，时间推移通常被称为自然康复现象。大多数有效的治疗都是由于治疗效果和安慰剂效应以某种不为人知的组合而产生的效果。正如多兹（Dodes, 1997）指出的："即使严重的疾病也有恶化和缓解的时候，关节炎和多发性硬化症就是典型的例子。"（Dodes, p.45）

大多数具有疗效的疗法都是有效治疗成分和安慰剂效应的未知组合。多兹（Dodes, 1997）同时也警告说，对于安慰剂的积极反应并不意味着病人的病是虚构出来的，他还警告，与流行的观念正相反，安慰剂可以是有害的："安慰剂效应能够通过证实或强化想象中的疾病来'诱发'慢性病。病人会对那些利用安慰剂效应的非科学从业者产

生依赖。”（Dodes, p. 45）

在关于心理治疗效果的研究中，怎样合理地对待安慰剂效应控制组，往往令人颇费周折。但是，这些复杂的问题不是我们在这里所要关注的，理解研究者为什么要将药物治疗的真实效果与安慰剂效应及自然康复区分开却很重要。

例如，研究表明，心理疗法确实优于只用安慰剂所产生的效果（Engel, 2008; Shadish & Baldwin, 2005）。但是，使用了安慰剂控制组的实验也表明，仅报告有多大比例的人感觉自己有所好转，会严重高估治疗的实际效果。问题就在于，得到见证叙述简直不费吹灰之力。康奈尔大学的心理学家托马斯·吉洛维奇（Thomas Gilovich, 1991）指出：“人类拥有如此容易自愈的身体，即使医生不做任何事情，很多寻求医学帮助的人也将体验到积极的疗效。如此一来，当自然康复的比率很高的时候，即使是毫无价值的治疗手段也能显得有效。”（p. 128）简而言之，无论干预的效果如何，只要运用治疗干预，潜在的安慰剂效应就会显现。问题在于，安慰剂效应是如此强大，以至于无论某个人使用的疗法多么荒唐，只要是被应用于一大群人的话，总有一些人会乐于为它的效果作出见证（清晨头部击打疗法，每天使用让你神清气爽！给我寄10.95美元，你就可以得到这个特制的、经过医学测试的橡胶锤）。

但我们确实不应该拿这种严肃的事情开玩笑。轻信见证叙述和个案研究的证据可能会导致灾难性的后果。回顾第2章，曾为抽动性秽语症作出科学界定——将之定义为器质性紊乱——的研究小组指出，人们对于个案研究证据的错误依赖，使得关于该病的、不可被证伪的精神分析理论长期盘踞不去，阻碍了对于该病病理进行真正的科学研究。发表在《新英格兰医学杂志》（New England Journal of Medicine）上的一篇社论，论述了在医学科学的从业者眼中个案研究和见证叙述的地位。“如果这本杂志收到一篇论文，说一个患胰腺癌的病人在服用了大黄根（rhubarb）后康复了……我们可能会发表一篇个案报告，但是，我们发表它并不是宣告它为一种新的疗法，而仅仅是推荐它作为一个值得用正规的临床实验进行验证的假设。与之相反，关于各类偏方秘方的轶闻（通常发表在通俗书籍和杂志上）则没有作出这样的声明，并且这些轶闻本身也不足以作为支持那些疗效的文献。”（Angell & Kassirer, 1998, pp. 839—840）

“鲜活性”问题

安慰剂效应的存在，宣告了见证叙述作为证据是无效的。这么做尽管很痛快，但是我们必须意识到，还存在着另外一个障碍，它阻碍了人们理解这一问题。社会和认知心理学家已经研究了人类记忆和决策中的所谓“鲜活性效应”。当面临问题解决或决策情境的时候，人们会从记忆中提取与当前情境有关的信息。

因此，人们倾向于利用更容易获得的、能够用来解决问题或做出决策的信息。对可获得性造成强烈影响的一个因素，就是信息的鲜活性。问题在于，再没有比发自内心的个人见证更鲜活、更引人注目的了，这都是一些已经发生的事或者是真实的事。个人见证的鲜活性常常令其他一些更可靠的信息黯然失色。购物时，我们在不同的品牌前权衡了半天，最后却由于某个朋友或某则广告对另一产品的推荐，而在最后一刻放弃了自己的选择。买车就是一个典型的例子。在翻看了《消费者报告》（Consumer Reports）中的数千份消费者调查之后，我们终于决定要购买一辆A品牌的车；又参考了几本汽车杂志之后，看到里面的专家们也都推荐A牌子的车，这更坚定了我们的选择。直到在一次聚会上，我们遇到一位朋友，他说他一个朋友的朋友买了一辆A牌子的车，结果是辆残次品，光维修就花了几百美元，而且这哥们决定再也不会买这个牌子的车了。显而易见，这样一个个案本不该在很大程度上影响到我们的决定，因为我们是在收集了针对数千名用户所做的调查报告和众位专家的评判之后才决定要买A牌子的车的。然而，我们中究竟有多少人能做到不把这个个案看得很重呢？购买汽车的例子说明，鲜活的个人见证所造成的问题并非心理学领域所独有。鲜活性会影响人们的决策，这样的例子无论在哪个领域都不难找到。作家迈克尔·刘易斯（Michael Lewis, 1997）描述了政治评论家乔治·威尔（George Will）——一个声名狼藉的反对政府干预的人，是如何在目睹了一场发生在其家门口、导致有人死亡的车祸之后，发表专栏文章呼吁强制使用安全气囊的。

设想一下，一个周五的早上，你在报纸上看到下面这样一个标题：“喷气式客机坠毁，413人死亡。”天啊，你也许会想，多可怕的事故啊！发生了多么糟糕的事情啊！继续设想，在接下来一周的周四，你起床看到报纸写道：“另外一场空难，442人死亡。”“哦，

不！”你也许会想，“不要再有任何灾难了，多么可怕啊，我们的空运系统怎么了？”然后想象一下——请尽可能地想象接下来的周五你起床时看到的是：“第三起空难悲剧：431人死亡。”不但是你，整个国家都会抓狂的。联邦政府会被要求尽快调查此事，所有航班禁飞，各种调查委员会如雨后春笋般成立起来，还有海量的法律诉讼被提起。《新闻周刊》和《时代》杂志将会对此作封面报道，它还会占据近期的电视新闻节目的头条。电视纪录片将会对此主题做深度挖掘。躁动和喧嚣是巨大而深远的。

这并不是一个虚构出来的问题，它是真实的。每周都有喷气式客机坠毁。也许不是一架巨型喷气式客机，而是很多小型飞机；或者也不是小飞机，而是小型交通工具，这种小型交通工具叫做汽车。在美国，每周都会有350多人死于汽车交通事故（每年超过19000人），人数足够坐满一架巨大的喷气式客机的了。

每周在高速公路上死于车祸的人数，相当于一架喷气客机的载员数，但我们对此漠然置之。这是因为“能坐满一架喷气式客机的人死了”这一信息没有通过媒体以一种鲜活的形式传达给我们。因此，每周死于汽车交通事故的350人（加上每周死于摩托车的80人），对我们来说不具有鲜活性。我们在餐桌前不会像谈论一架喷气式飞机坠毁并且死了很多人那样谈论这些死于车祸的人，我们不会就汽车出行的安全性和必要性进行争论。但是，如果大型喷气式客机每周都发生坠毁，并且每次都导致350人死亡的话，我们就会讨论空运交通的安全性。车祸中死亡的这350人不会上新闻，因为他们分布在全国各地，因此对于我们中的大多数人来说只是统计学上的抽象概念。媒体不会为我们生动地呈现这350名死者，因为他们并不是死在同一个地方。相反，媒体呈现给我们的（有时候）是一个数据（例如，每周350人）。这已经足够引起我们的思考了，但是我们对此毫无反应。与我们生活中的其他任何行为相比，驾驶汽车都是一种极端危险的行为（Galovski, Malta, & Blanchard, 2006; Gardner, 2008; National Safety Counil, 2001）。然而，关于它的风险和相对应的收益，从未有过全国性的大讨论。这对于住在郊外、需要驾车往返的人来说，是不是一个可以接受的代价？我们从不去问这样的问题，因为问题还没被意识到，原因就是：代价和风险没有像空难那样以鲜活的方式呈现给我们。

想想下面这个例子的荒谬之处吧。一个朋友开车20公里载你去机场，因为你要乘飞机作一次750公里的旅行。分别的时候，你的朋友很可能会说：“一路平安”，这个临别赠言其实是充满伤感的讽刺意味的，因为你的朋友在回家的20公里路上死于车祸的风险，要比你飞行

750公里的风险高出3倍。这就是鲜活性问题，它解释了A对B的安全祝福存在着明显的不合理性，因为恰恰是A正处在更大的风险之中（Sivak & Flannagan, 2003）。这些例子并非只是假设，在“9·11”恐怖袭击事件之后，乘飞机出行的人数锐减，因为人们害怕飞行。当然，人们还要继续外出旅游，而不只是待在家里。他们只是改为其他方式出游——大多数情况下都是自驾车。但是，自驾游比飞行要危险得多，从统计学上讲，注定有更多人因转成自驾游而死亡。事实上，研究者估计，在2001年的最后一个月，有300多人由于乘坐汽车而非飞机旅行导致死亡（Gigerenzer, 2004, 2006）。有一个研究团队能够以一种鲜活的统计来传达出驾驶机动车有多么危险。西瓦克和福兰纳根（Sivak & Flanagan, 2003）计算出，如果驾车和乘坐飞机的危险系数是一样的话，那么“9·11”这个级别的事故将会每个月都发生一次！

这就是为什么尽管飞机上会推荐使用婴幼儿专用座椅（具备核准的束缚紧度），联邦航空管理局（FAA）也不对其进行强制要求（美联社，2010）。他们不会这么要求的原因是FAA担心如果强迫父母为婴幼儿买座位，很多父母就会选择驾车而不是坐飞机——当驾车时孩子在父母的腿上时，它们比在飞机上要危险得多。在我们日常生活环境下，没有什么地方比让孩子呆在车里更危险，许多父母只是认识不到这个事实。

我们很难在判断中避免鲜活性的影响。比如说康奈尔大学，因其学生自杀率高而出名。我们必须问一下为什么它有这个名声。之所以这么问，是因为从统计上讲，它并不是自杀率高的学校。事实上，它的自杀率在全美处于中下水平（Frank, 2007）。它的名声和实际的统计数据没有一点关系——与康奈尔的实际自杀频数一点关系都没有。与之有关的事实是，康奈尔大学挨着两道冰封的深渊峡谷——横跨峡谷的是两条令人胆战心惊的桥梁（Frank, 2007）。无怪乎自杀案件经常在这些桥上发生，当救援队在峡谷里找回尸体时，往往会导致交通瘫痪，更重要的是，生动电视画面拍摄到了自杀案的一组组镜头。而服药过度导致的死亡则没有等量齐观的媒体报道。可见康奈尔大学的高自杀率名声来源于生动性，而不是统计数据。

在媒体的帮衬下，鲜活性误导个人判断的情况也同样广泛存在于其他领域里。一项研究（Cardner, 2008; Radford, 2005; Skenazy, 2009）调查了父母最担心他们的孩子遭遇哪种风险。结果显示，父母最担心的是孩子遭绑架，而这一事件发生的概率是1/600000。相比之下，父母则不太担心孩子在车祸中身亡的危险，然而这种可能性比遭绑架要高出几十倍（Gardner, 2008）。同样地，儿童在游泳池里

溺死的概率要比他们被绑架和被陌生人杀死的概率高得多（Kalb & White, 2010）显然，对绑架的担心大部分是由于媒体渲染的结果。车祸、意外事故（包括枪支意外）、儿童期肥胖、青少年时期自杀可能比绑架和鲨鱼袭击对我们的孩子的幸福威胁更大，正如像科学作家加德纳（Dan Gardner, 2008）写的那样："我们总是易受恐怖场景的伤害。"（p.84）举例来说，这类"恐怖场景"导致了父母年年担心万圣夜的有毒糖果，然而实际上从来没有一个记录在案的事件说有任何孩子在万圣夜因为了吃毒糖果而死亡（Skenazy, 2010）。

媒体制造的鲜活性效应，使得我们的风险认知发生了紊乱。例如，不断增长的糖尿病发病率导致了人们对其风险的低估，甚至低于对住院会增加感染葡萄球菌的风险的担心，然而前者每年影响4500万美国人，而后者仅影响1500人（Fountain, 2006）。尽管就个人而言，我们对糖尿病可以做很多事（改变饮食习惯和锻炼），而对后者相对无能为力。

呈现的鲜活性甚至可以影响我们对科学证据自身的解释。在一项研究中，被试看到一些针对心理现象的描述和解释（Weisberg, Keil, Goodstein, Rawson, & Gray, 2008）。其中一些是好的解释（包括真实的心理概念），另一些是糟糕的解释（只是用循环的语式对现象进行循环描述，而不是解释它）。当这些解释前加了一句"大脑扫描显示"时，对两类解释质量的评价（尤其是糟糕的解释）显著地上升了。相似地，麦凯布和卡斯特尔（McCabe & Castel, 2008）发现，在认知神经科学领域，如果实验结论中包含概括结果的脑成像图片，人们对这一结果可信程度的评价要高于描述相同结果的图表。简而言之，对科学结果呈现的鲜活性也会影响对研究的评价。

单一个案的压倒性影响

一个很著名的关于人们如何对鲜活的轶事信息作出不同反应的例子，来自于在20世纪60年代中后期媒体对越战的报道。随着战事的拖延，美军的死亡人数仿佛无休止地增加，媒体开始报道当周美军死亡的人数。一周接一周地过去了，这个数字在200至300之间徘徊，公众似乎已对这种报道习以为常了。

然而，某杂志用几个版面的篇幅连续刊登了前一周阵亡者的个人照片。这时，公众非常具体地看到了在这样一个有代表性的一周内逝去的大约250个鲜活的生命。结果，此举导致了大规模的、针对这场代价巨大的战争的抗议声浪。250张照片所产生的影响是每周数字报

道所远不能及的。但是作为一个社会成员，我们应该克服这种不相信数字、必须眼见为实的倾向。绝大多数影响我们社会的复杂因素都只有靠数字才能捕捉。只有当公众学会像重视图像材料一样重视以数字形式表达的抽象材料时，公众自己的立场才不会像屏幕上闪过的最新图像那样变化无常。

2004年，一档叫做《晚间在线》的电视节目在伊拉克战争一周年之际公布了在这场战争中死亡的700多名战士的名字和照片。在这一时刻，历史又重演了。这一做法与该档节目在“9·11”事件一周年之际播放受害者的姓名和照片的套路完全相同，当时这些照片的播放都征得了受害者家属同意。

然而，死亡士兵的照片还是引发了战争支持者的抗议。有些人控诉节目主持人泰德·考佩尔对这场战事抱有敌意，但是这些指控显然瞄错了对象，因为考佩尔并不反对这场战争（CNN.com, 2004）。实际上，战死的人数并非没有被报道，这700多人死亡的消息日复一日地出现在这个国家的每一份报纸上。但是争论的双方都知道，公众尚未对那些数字进行“加工”——没有计算代价，是因为那些数字还过于抽象。双方都知道很多人在看过这些照片之后，都会从头对这些信息进行加工，并开始真正在意战争的代价。

不仅公众受到鲜活性问题的困扰，在心理学和医学领域，有经验的临床从业者一直都在努力摆脱个别案例的压倒性影响给他们的决策带来的阴影。作家弗兰辛·卢索（Russo, 1999）描述了弗吉尼亚大学的肿瘤专家威利·安德森（Willie Anderson）面临的两难境地，安德森一直提倡控制实验，并会定期招募一些病人来做有控制的临床测试。但是他仍旧纠缠于自己对突出个案的反应，那些鲜活的个案对他的决策产生了影响。尽管他相信科学，但仍承认“当真实的人眼巴巴地看着你的时候，你将被他们的期望以及自己对他们期望的期望所包围，这确实非常困难”（p. 36）。但是安德森知道，有时对他的病人来说，最好的办法就是忽略“看着你的那个真实的人”，并且遵循最佳证据的指示。最佳证据来自于有控制的临床试验（将在第6章表述），而不是看着你的那个人的情感反应。

总之，过于依赖见证证据的问题一直存在。此类证据的鲜活性常常掩盖了更加可靠的信息，并且混淆视听。心理学教师担心的是，仅仅指出依赖见证证据的逻辑谬误，并不足以让人们从一个更深的层次理解这类数据的缺陷。

我们还能做些什么呢？还有什么其他的方法能让人们理解这个概念吗？幸运的是，我们还有一个法宝——一种与学术方法略有不同的

方法。这种方法的本质是以鲜活性来对付鲜活性，是以一种“以彼之道，还施彼身”的方法对付见证证据，让见证用自身的荒谬来击溃自己。这个方法的实践者，就是独一无二、毋庸置疑的“了不起的兰迪”！

了不起的兰迪：以彼之道，还施彼身

詹姆斯·兰迪（James Randi）是一位魔术师，并且是个多面手，他曾经被麦克阿瑟基金会授予过“天才”奖。多年来，他一直尝试着教公众学会一些基本的批判性思维的技巧。“了不起的兰迪”（Amazing Randi, 他的艺名）通过揭穿“通灵”骗术和庸医疗法来达到教育公众的目的。尽管他拆穿了很多魔术和伪装的所谓“通灵术”，但最为著名的还是他拆穿20世纪70年代通灵术超级明星尤里·盖勒（Uri Geller）的把戏的那一回。盖勒靠吹嘘通灵术红透荧屏，他对媒体的蛊惑程度简直可以用无以复加来形容。各大洲的报纸、电视节目和主要的新闻杂志对他争相报道（盖勒仍健在，还在写书；Radford, 2006）。兰迪发现并揭露了盖勒经常表演的通灵术“绝活”其实不过是些普通和简单得令人难以置信的魔术把戏，包括使勺子和钥匙弯曲、使钟表开始走动，等等，这些对于一个优秀的魔术师来说简直就是家常便饭。自从盖勒被拆穿以后，兰迪继续将他那非凡的才智用于维护公众的知情权，他不断揭露超感官感知、生物节律、超自然力、通灵外科手术、天外来客、漂浮术以及其他伪科学的谬误，以帮助公众了解真相（Randi, 1995, 2005, 2011; Sagan, 1996; Shermer, 2011）。

兰迪的另外一个兴趣就是去证明，对于任何一个荒谬的事件或无中生有的言论而言，获得见证是多么容易。他的手法就是，让人们掉进其见证所编织的陷阱里。在一档广播节目中，兰迪揭示了另外一种伪科学——生物节律能够如此流行的原因（Hines, 1998, 2003）。一位听众同意每天都记日记，并将日记与一份特别为她准备的两个月的生物节律表做比较。两个月以后，她打回电话告诉听众：生物节律绝对不是假的，因为节律表预测实际行为的准确率超过了90%。兰迪不得不把他的秘书所犯的一个愚蠢的错误告诉给这位听众，秘书错误地将本该发送给另外一个人的节律表发给了她，而不是她自己的。然而，这位妇女还是同意看一下真正属于自己的表格是怎样的，于是，又一份表格立即发送给了这位妇女，并且请她再打电话过来。几天后，这位妇女带着解脱感打进电话，说她自己的表格也同样十分准确——事实上，更为准确。在下一期节目中，大家发现，另一个错误发生了。这位妇女收到的是兰迪秘书的节律表，而不是她自己的！

兰迪的生物节律和占星术小把戏，其实是一种被命名为巴纳姆效应（Barnum, 著名的嘉年华和马戏团的团主，提出了“每分钟都会有人上当受骗”的说法）现象的范例。这一效应曾被心理学家广泛地研究。研究者发现，大多数成年人都会认为泛化的个性总结都是准确的，并且都是对自己独特的描述。这里有一个来自谢尔默的例子（2005, p.6）：

> 你是一个非常体贴的人，总是及时地帮助别人。但是也有一些时候，你会发现你有一点点自私……有时候你太忠于自己的感受以至于会暴露过多的自己。你善于思考，并且对任何事情，在改变想法之前都希望看到证据。如果你处在一个陌生的环境下，你会非常小心，直到你看清楚发生了什么事情，然后才会充满信心地行动……你知道怎样做一个好朋友，你懂得训练自己，所以在别人看来，都在你的一切掌控之中，但其实有些时候你是缺少安全感的。你希望在人际关系中比现在更受欢迎，更加自如。你面对世界表现得很有智慧，这种智慧来源于艰难的体验而非书本学习。

大多数人都发现，这个总结是对其个性非常准确的概括，但是很少有人自发地意识到大多数其他人也同样认为它描述了他们自己！许多众所周知的语句和措辞（如这个例子）使很多人认为适用于他们自己。谁都能够将其作为一个个人化的心理“分析”提供给“顾客”，而这些顾客常常会为个人化的“性格解读”的“准确性”而感到震惊，却不知道其实每个人的解读都是一样的。

当然，巴纳姆效应正是手相学和占星术的基础（Kelly, 1997, 1998）。巴纳姆效应还可以证明产生见证有多么容易，以及为何见证毫无价值。这就是詹姆斯·兰迪运用这些小把戏努力想要达到的目的——给人们好好上一课，告诉人们见证证据是没有价值的。他不断地证实，形成有利于虚假主张的见证是多么容易。正是这个原因，用见证来支持自己提出来的特定理论是毫无意义的。检验一个主张时，只有来自有控制的观察中的证据（第6章中将会描述）才是足够充分的。

见证为伪科学打开方便之门

有时候有人会说，类似刚才所讨论的种种伪科学，只不过是人们给自己找乐子的一种方式，无伤大雅。再者说，我们又何必较真呢？不就是有几个人在异想天开，而另外几个人从中赚点儿小钱吗？实际上，对此问题进行一番彻底的考察就不难发现：伪科学的盛行对社会的危害比人们想象的要严重得多。

首先，人们倾向于不考虑经济学家所说的“机会成本”。如果你花时间做一件事，你已经失去了做另一件事的时间。你也失去了花费时间的其他机会。当你在一件事上花费了金钱，你就失去了花钱做其他事的机会——你失去了让钱花在其他地方的机会。伪科学存在大量的机会成本。当在伪科学上花费了时间（和金钱），人们不但没有收获，还浪费了本可以花在更有价值的事情上面的时间。

对此问题进行一番彻底的考察，就不难发现伪科学的盛行对社会的危害比人们想象得要大得多，并且其花费超过了机会成本。在一个复杂的、科技化的社会中，一些能够影响千万人的决策会为伪科学的影响推波助澜。也就是说，即使你并不认同这些伪科学的观念，你也可能受到这些观念的影响。

例如，三分之一的美国人都喝无氟的水，尽管大量的科学证据表明含氟的水可以显著减少蛀牙（Beck, 2008; Griffin, Regnier, Griffin, & Huntley, 2007; Singh, Spencer, & Brennan, 2007）。疾病控制中心估计，在氟上每花费1美元，在牙病治疗上的花销会节省38美元（Brody, 2012）。无氟地区数以百万的美国人正承受着不必要的蛀牙的折磨，仅仅因为他们的邻居坚持相信伪科学的阴谋论——氟有很多有害的影响。一小撮怀揣这类伪科学观点的人已经让许多社区远离氟，并且对周围的每个人渲扬氟的坏处。简而言之，少部分人的伪科学观点会使大多数人受到消极影响。

再看这样一个例子，即使你不相信伪科学，你也可能被它影响。大银行和一些500强企业雇用笔迹学家来做人事选拔，即便大量的证据表明，笔迹学在实现这一目的方面是没有作用的（Lilienfeld et al., 2010）。伪科学的笔迹学指标在一定程度上使雇主忽视了其他更有效

的选拔标准，导致的结果是经济上的零效益和对一些人的不公平待遇。如果仅仅因为笔迹中有一个连写的“小圈圈”，就让你丧失了获得一份很心仪的工作的机会，你会作何感受？

不幸的是，这样的例子绝非凤毛麟角（Shermer, 2005; Stanovich, 2004）。当伪科学的观念渗透于整个社会的时候，我们都以不同的方式受到影响——即使我们并不认同这些信念。例如，警局雇通灵师协助办案，即便研究表明这一举动是没有任何效果的（Radford, 2010; Shaffer & Jadwiszczok, 2010）。没有一个记录在案的案例表明通过通灵信息能够成功地找到失踪的人（Radford, 2009）。

如今，类似占星术这样的伪科学是一项巨大的产业，涉及报纸专栏、广播节目、图书出版、网络、杂志文章以及其他各种传播渠道。星相学杂志的发行量要比很多正规的科学杂志大得多。据美国众议院老龄化问题委员会估算，浪费在医疗骗术上的钱已经达到数十亿美元。简而言之，伪科学是个油水颇丰的行当，数以千计的人靠公众的盲信盲从而获得收益。

在抨击伪科学方面，一些协会和组织比心理学更激进。2007年，美国联邦交易委员会（FTC）对通过电视宣传和名人代言销售减肥药物的四个商家处于了数百万美元的罚金。在宣布罚款时，FTC女主席狄波拉·普拉特·梅杰拉斯（Deborah Platt Majoras）试图教育大众道：“美国人需要明白，个人见证不能代替科学。”（de la Cruz, 2007, p. A10）与之相似，医学界的各类组织都比心理学界表现得更为激进和勇猛。下面就让我们看看由关节炎基金会出版、曾被美国众议院老龄化问题委员会所引述的一套识别不道德药品推销员的指南。

1. 他或许会提供一种用于治疗关节炎的“特别的”或“秘密的”处方或设备。

2. 他会做广告，用的都是“个案史”和“满意患者”的见证。

3. 他或许会承诺（或者暗示）能够快速或轻松见效。

4. 他也许会声称知道关节炎的成因，并且说能够“清除”你体内的“毒素”，同时促进你的健康。他或许会说外科手术、X光和医师所开的处方是没有必要的。

5. 他或许会指责“医学体制”故意阻碍了进步，或者迫害了他……但是他不允许他的方法以已有的或已获证明的方法来验

证。

这份清单同样可以作为识别带有欺骗性的心理学疗法和理论的指南。在这里，请注意上面第2条，这正是本章关注的焦点。同时注意，第1条和第5条论证了之前所讨论过的一个观点：科学是公开的。除了宣扬见证叙述作为“证据”，伪科学的从业人员经常以指责他人有意要压制他们所获取的“知识”，来试图逃避“公开可证实”这一科学的标准。这样，他们就有借口带着他们的“研究成果”直接走进媒体，而不是通过正规的科学出版程序将这些成果公诸世人。

还可以加入上述清单里的一个警告是，小心提防这种情况：有人似乎是正在兜售某些东西，能够让人脱离已经确立的一些自均衡。例如在投资领域，众所周知，风险和回报是密切相关的（回报越高意味风险越高）；在减肥领域，人人都知道长期的体重下降依靠的是卡路里摄入量的长期改变；至于教育干预，众所周知的是，持久的教学效果需要长期的高密度的干预程序。总之高风险高回报自均衡；体重下降和卡路里摄入自均衡；学习提高和干预强度自均衡。在这些方面，推崇伪科学的人总是声称他们能够打破这些自均衡——没有风险，你也有高回报；你使劲吃也能减肥；短期的干预就能显著地改善你的学习成绩。你可以确信那些违反基本自均衡的主张是虚假的。例如，你可以确信那些叫“爱婴斯坦”（Baby Einstein）的产品绝不会具有与它名字所暗示的相同功效（Broson & Merryman, 2009; Deloache et al., 2010）。

重要的是，我们要意识到电视、网络、平面媒体会公布心理学领域里几乎所有稀奇古怪的主张，只要他们觉得这能够吸引观众，就不顾这些主张与已有的证据是多么地相悖。媒体呈现的是一些似是而非的言论和专家——混合了真正的科学家和伪科学江湖骗子。让我们重新回顾一下在电视上流行了20年的奥普拉脱口秀。公平地讲，必须得说节目经常邀请一些可信的专业人士，传达给观众一些从乳腺癌到个人理财等许多领域的重要信息。但是，混入到节目中的还有真假莫辨的最无耻的江湖骗子（Gardner, 2010）。例如，奥普拉宣扬某人使用塔罗牌来诊断疾病的另类疗法以及某人认为女性的甲状腺问题是由于“一辈子想说的话说不出口”导致“能量滞留在了喉咙区域”（Kosova & Wingert, 2009, p.59）。

几年前，《奥普拉脱口秀》热衷于传播称之为“秘密”的一大堆心理呓语，其默认前提是“所有的疾病都能够由心念力量所单独治愈”（p.61）。但是在这些特定的节目播出后，暴露真相的事情发生了。一个叫金·汀克汉姆（Kim Tinkham）的妇女给奥普拉写信说，她得了乳腺癌，但是不打算接受手术和化疗，不仅她的医生建议她化

疗，她所收到的第二和第三建议也是化疗。金告诉奥普拉，她不打算手术，她要遵从“秘密”。奥普拉看到这封信感到很震惊，于是她让这个妇女参加节目并且极力劝她接受手术和化疗。金·汀克汉姆现在已经去世，但当奥普拉恳求她时，人们已经了解了在她节目上“秘密”的错误——“我只是说它是一个工具，它不是所有事情的答案”（p.62）。不，奥普拉，这不是工具，这是伪科学。当你像这样在节目中把科学和伪科学混杂在一起时，即使是出于好意，像金·汀克汉姆这类受到伤害的例子总会出现。

正如这个悲剧例子所显示的，由于陷入伪科学之中，人们无法利用对他们真正有效的治疗手段。许多病人之所延误了接受正确的医学治疗，就是因为他们浪费时间去寻求虚假的治疗。史蒂夫·乔布斯在被告知得了胰腺癌后，就无视他的医生而去追求未经证实的水果节食、咨询灵媒、进行无效的水疗，延误手术长达9个月之久（Isaacson, 2011）。

最后，看一下10岁的坎迪斯·纽梅克（Candace Newmaker）的悲伤例子。由于违纪问题，她的养母把她带到一个叫作“儿童依恋治疗及训练协会”的地方（Shermer, 2012）。根据疗法背后的伪理论，某些儿童需要“面对”和“抑制”，能够让他们走出所谓的“压抑了的被遗弃的愤怒”。坎迪斯被床单和枕头蒙住脑袋，同时成年人压在她的头顶上以便于她能够“重生”。当坎迪斯大哭大叫时，成年人压得更使劲并且大呼坎迪斯为“懦夫”。在40分钟的这种荒谬的行为之后，坎迪斯安静了。她死了。死于窒息。

社会总是对这些危害人类的伪科学家们太过温和，但这次不是。她的所有治疗师由于不计后果地虐待儿童导致儿童死亡而被判了16年监禁。迈克尔·舍默（Michael Shermer, 2011）评论道，尽管验尸显示，孩子死于“缺氧缺血性脑病变导致的脑水肿和脑疝，但终极原因是伪科学的骗子伪装成心理学家……这些治疗师杀死了坎迪斯，并不是因为他们是邪恶的，而是因为他们受到了以迷信和巫术思维为基础的伪心理学信仰的控制”（p.86）。

通过一个例子能够清楚地看出，当伪科学观念广泛传播时，我们大家是如何受到伤害的。有个理论（首先在20世纪90年代初被提出，迄今仍然在流传）认为孤独症与儿童时期的疫苗接种有关。这个理论是错误的（Grant, 2012; Honda, Shimizu, & Rutter, 2005; Judelsohn, 2007; Novella, 2007; Offit, 2008; Taylor, 2006），但是本章的读者应该不会对这一观念是如何产生的感到惊讶。

很多儿童在第一次接种疫苗前后被确诊为孤独症，并且其中许多

开始出现明显可辨的症状（语言学习延迟、社会交往困难、受限制的行为模式）。毫不奇怪，假如有数以千计的儿童有这样的状况，总有一些父母在儿童接种疫苗之后很短的时间内注意到儿童的这些症状（通过诊断或基于自己的观察而逐渐意识到）。随后这些家长会提供生动而真诚的见证，证明他们孩子的状况一定和接种疫苗有关。然而，许多各类实验和流行病学研究都得出了聚合性的结论（见第8章）：这种关系是不存在的（Deer, 2011）。这一伪科学的信念会让涉入的家长和孩子有更多的损失，而不仅仅是机会成本。这个错误的观念引发了一场反疫苗运动，结果就是免疫率下降，导致更多的儿童因麻疹住院，有些甚至死去，而这原本可能不会发生（Goldacre, 2008; Grant, 2011; Judelsohn, 2007; Novella, 2007; Offit, 2008）。这再次证明，在一个相互联系的社会，你邻居的伪科学观念可能会影响你，即使你不相信这样的观念。

当政治领袖笃信伪科学时，他们的信仰会给数以千计的人们造成严重的后果。南非前总统塔博·姆贝基（Thabo Mbeki）拒绝接受“艾滋病是由病毒引起的”这一科学的共识（Pigliucci, 2010）。邻近的国家博茨瓦纳和纳米比亚给感染艾滋病的民众注射抗逆转录病毒，但是南非不这样做。据估计，拒绝接受抗逆转录病毒导致365000名南非人遭受过早死亡（Singer, 2008）。

心理学家越来越关注医学骗局在互联网上的蔓延（Offit, 2008），以及它对健康带来的损害。麦克斯·考皮斯（Max Coppes）博士不得不给《新英格兰医学杂志》（New England Journal of Medicine）写了一封信，警告人们注意医学中的伪科学所带来的危害（Scott, 1999）。他描述了一个9岁女孩的案例，这个孩子在经历癌症手术之后，如果接受化疗的话，将会有50%的机会可以多活3年。但她的父母找到一种未经验证的、利用鲨鱼软骨的偏方来代替化疗。小女孩在4个月后就失去了生命。

当我正在讲述这个话题的时候，经常有人会针对我的演讲提出非常中肯的问题：“你不也是正在用生动的个案来阐述你的观点吗？这种做法难道不正是你反对的吗？”这个问题问得好，并且它让我有机会详细阐述本章中包含的一些论点间的微妙之处。这个问题的答案是肯定的，我运用了生动例子来阐述观点。但是，这是为了阐述观点，而不是为了证明观点。这里的关键是要区分两点：主张的提出和主张的交流。对于每个主张，我们都能问这样一个问题：它是不是基于鲜活的见证？这会产生四种可能的情况：

1. 一项主张基于鲜活的见证，同时依靠鲜活的见证来交流；

2. 一项主张基于鲜活的见证，同时不依靠鲜活的见证来交流；

3. 一项主张基于证据而非鲜活的见证，同时依靠鲜活的见证来交流；

4. 一项主张基于证据而非鲜活的见证，同时不依靠鲜活的见证来交流。

本章中的一些讨论属于第3种情况：一项主张基于证据而非鲜活的见证，同时依靠鲜活的见证来交流。例如，我引用了很多非见证的证据贯穿整章，就是为了说明，个案研究的证据不能用于建立因果性结论，鲜活的例子在人们的判断中被赋予了过高的权重，伪科学的代价巨大，等等。对于这些主张中的每一项，我都标出了引证和参考文献。尽管如此，出于交流的目的，我使用了一些鲜活的案例，将注意力吸引到这些主张上，并让它们给人们留下深刻的印象。关键的一点是，支持这些主张本身的并不仅仅是鲜活的见证。比如，我曾使用一些鲜活的例子来阐述“鲜活的例子在人们的判断中被赋予了过高的权重”这一事实，但是这一主张的证据包含在我所引用的经过了同行评议的科学证据之中（例如，Li & Chapman, 2009; Obrecht, Chapman, & Gelman, 2009; Sinaceur, Heath, & Cole, 2005; Slovic, 2007; Wang, 2009）。

回到这部分的主要观点上，并做个总结吧，那就是伪科学的传播所造成的代价是巨大的。需要搞清楚哪种类型的证据能够揭示某种现象中蕴涵的道理或理论是否可信，如果搞不清楚这一点，就会大大有利于伪科学的传播。由于见证叙述可以为任何主张提供唾手可得的支持，以及自身所具备的冲击力，见证打开了通往伪科学的大门。对于心理学信息的消费者来说，对它们保持警惕应当是头等大事。在接下来的几章中我们将会看到，在证实某种主张的合理性时，究竟需要哪些类型的证据。

小结

个案研究和见证叙述在心理学（以及其他科学）研究的早期阶段是有用的，因为此时寻找有趣的现象和待研究的关键变量很重要。虽然个案研究在早期的、理论形成前的阶段是有用的，但在研究的后期，当对理论进行检验之时，个案研究就毫无用处了。这是因为，作为一个孤立现象，个案研究的结果遗漏了太多其他可能的解释。为何个案研究和见证证据对于理论检验来说是没有用的？要想理解这一点，就需要想一想安慰剂效应。安慰剂效应是指无论疗法是否包含了有效的成分，人们都倾向于报告任何疗法都对他们有效。安慰剂效应的存在，催生了许多关于疗效的见证叙述，致使对一种心理（或医学）疗法效果的证明成为“不可能的任务”。原因就在于，无论治疗手段是什么，安慰剂效应都会使人们提出证实其疗效的个人见证。

尽管见证证据在检验理论的时候是无用的，但心理学研究指出，由于鲜活性效应，这类证据经常被人们过分地倚重：对于更为生动并因此在记忆中更易提取的证据，人们会赋予其过高的权重。对大多数人来说，见证证据就是一种格外生动和鲜活的信息。因此，人们在验证某一心理学主张的合理性时，会过度依赖这类证据。事实上，理论主张是否合理，是不能用见证叙述和个案研究的证据来判定的。

Chapter 5
相关和因果：用“烤箱法”避孕

多年前，在中国台湾地区曾开展过一次大规模的研究，目的是调查哪些因素是与人们对避孕工具的使用有关的。一个由社会学家和内科医生组成的大型研究团队收集了有关环境和行为变量方面的大量数据。研究者比较感兴趣的是，哪种变量能够最准确地预测避孕方法。数据收集上来之后，研究者发现，有一个变量和使用避孕工具的相关最强，这就是：家庭中家用电器（烤箱、风扇等）的数量（Li, 1975）。

这个结果恐怕不会促使你提出这样的建议：在高中发放免费的烤箱以解决青少年的怀孕问题。但是，你为何不会有这样的想法呢？电器和避孕工具使用之间的相关性很高，在众多被测量的变量中，这个变量是唯一最准确的预测因子。我希望你的回答会是：问题关键在于这两个变量间关系的“性质”，而非“强度”。开展“免费烤箱计划”预示着这样一种观念：烤箱导致人们使用避孕工具。而实际上我们会将这种建议视为一种荒唐的方案，至少在上面所举的这个显而易见的例子中，我们会认识到，这两个变量可能相关，但不是因果关系。

在这个例子中，我们可以猜想，这种关系之所以存在，是因为“避孕工具的使用”和“家庭中家用电器的数量”这两个变量通过与这两种变量都相关的其他变量联系起来。社会经济地位（SES）可能会是中介变量之一。我们知道，社会经济地位和避孕工具使用有关。现在我们所需要的就是这样一个事实：经济水平高的家庭会拥有更多的家用电器，我们都会有这样的联想。当然，其他的变量也可能会在二者的关系中起到中介作用。但是，无论“家用电器的数量”和“避孕工具使用”之间的相关有多么强，这种关系都不能说明它们之间存在因果

关系。

避孕方法的例子很容易让我们理解这一章的主旨：有相关，并不意味着必然有因果关系。在本章中，我们将会讨论阻止我们做出因果推论的两大问题：第三变量问题和方向性问题。我们还将会讨论选择性偏见是如何导致第三变量问题的。

相关性证据的局限性并不都像“烤箱”例子那样容易被识别。当因果关系对我们来说显而易见时，当我们抱有根深蒂固的偏见时，或者当我们的解释被理论定势所主宰时，就会很容易地把相关当作因果的证据。

第三变量问题：戈德伯格与糙皮病

在20世纪初期，数以万计的美国南部民众罹患并死于一种叫做糙皮病的疾病。糙皮病被认为是由一种不明微生物引发的传染性疾病，其主要症状是头晕、嗜睡、溃疡、呕吐和严重腹泻。此后，许多来自全国糙皮病研究学会的医生都认同这样的证据：糙皮病和卫生条件有关。这并不令人吃惊。家在南卡罗来纳州斯帕坦堡的人们似乎总是远离糙皮病的困扰，因为他们有自来水管道和良好的污水处理设施。相反，糙皮病患者家中的污水处理条件都比较差。这种相关恰好验证了这样的观点：由于糟糕的卫生条件，传染性疾病是通过糙皮病患者的排泄物传播开来的。

一名叫约瑟夫·戈德伯格（Joseph Goldberger）的医生对这种解释非常怀疑，在美国公共卫生部部长的指示下，戈德伯格针对糙皮病开展了许多研究。他认为糙皮病是由于营养不均衡的饮食引起的，简而言之，是美国南部普遍的贫困造成的。许多患者赖以生存的都是高碳水化合物、低蛋白质含量的饮食，如很少量的肉类、蛋类、牛奶，以及大量的谷类、燕麦和玉米粥。戈德伯格认为，污水处理条件和糙皮病之间的相关在任何一个方面都无法反映因果关系（和烤箱控制生育的例子一样）。他认为根本原因在于，拥有清洁管道的家庭通常也都是经济状况良好的家庭，经济上的差异也会反映在他们的饮食上，经济状况好的家庭在其饮食中包含更多的动物蛋白。

但是，等一下！为什么戈德伯格的因果推断就一定是对的呢？毕竟，两派人马都是坐在那里，根据相关数据推论什么才是造成糙皮病的原因的。为什么医学会的医生们不能说戈德伯格的相关同样也是误导性的呢？为什么戈德伯格能够推翻别人的假设——一种微生物通过糙皮病患者的排泄物传播，而这种传播是因为不完善的污水处理设施造成的？戈德伯格对糙皮病的判断还涉及一个小细节，这个细节我刚才没说：戈德伯格吃下了糙皮病患者的排泄物。

为什么戈德伯格的证据更好

戈德伯格有一类这样得来的证据：研究者不仅观察相关性，还靠真正地操纵关键变量来收集数据（有关控制操纵，将在第6章进一步讨论）。这种方法经常要创造一些通常极少会自然出现的条件——说戈德伯格设计的特殊条件不会自然出现，无论怎样强调都不会过分。

戈德伯格确信糙皮病是不会传染的，也不会通过患者的体液传播。他给自己注射了一名患者的血液，还吃进一名患者喉咙和鼻腔内的分泌物。据记载，戈德伯格、戈德伯格的助手以及戈德伯格的妻子自愿服下包含糙皮病患者尿液和排泄物的小面团（Bronfenbrenner & Mahoney, 1975）！不管这手段有多极端，结果是，无论是戈德伯格还是其他的志愿者都没有患上糙皮病。简而言之，戈德伯格创造了这个传染疾病可能传播的所有条件，结果平安无事。

戈德伯格对其他人提出的因果机制进行了操作，结果显示该机制是无效的。尽管如此，对他自己提出的因果机制进行检验仍然非常必要。戈德伯格选择了来自密西西比州监狱农场的两组犯人，这些人都是没有患糙皮病的，并且都是自愿参加实验。其中的一组人被给予高碳水化合物、低蛋白质的食物，这种类型的食物是戈德伯格怀疑引起糙皮病的原因。另一组被试被给予（营养成分）更均衡的饮食。5个月后，低蛋白质的这一组患上了糙皮病，而另一组却没有丝毫的患病迹象。戈德伯格的理论遭到了一些人的反对，这些人出于政治动机而否认贫困的存在。经过长期的抗争，戈德伯格的假设终于被人们所接受，因为他的假设与实验证据的契合程度是其他任何假设所不能比拟的。

糙皮病的历史说明，如果依据相关研究来制定社会和经济政策，必将使人类付出惨痛的代价。但这并不意味着我们永远不要使用相关研究的证据。恰恰相反，在许多场合，我们必须用到相关（见第8章），而在某些情况下，只要有相关就够了（例如，当我们的目标是预测而不是决定原因的时候）。科学家们经常不得不使用不充分的知识来解决问题。重要的是，我们在运用相关性证据的时候要谨慎小心。像“糙皮病—污水”这样的案例，在心理学研究的每个领域内都频频发生。这个例子也揭示了“第三变量问题”：事实上，两个变量之间的相关——这个例子中是糙皮病的发病率和污水处理条件——并不意味着这两个变量之间有直接的因果关系，相关之所以产生，是因为这两个变量都分别与第三变量相关——这里是饮食——而这个变量没有被测量。糙皮病和社会经济地位（还有饮食——真正的变量）相关，社会经济地位又和污水处理条件有关。像这种污水处理条件和糙皮病之间的相关，我们通常称之为“虚假相关”：相关的产生不是因为两个变量之间存在一个可以测量的直接的因果联系，而是因为这两个变量

都与第三变量相关。

下面我们来看一个发生在现实生活中的例子。多年以来，有关公立学校和私立学校教学质量的争论甚嚣尘上。从这场争论中得出的一些结论，很生动地展示了从相关证据推出因果关系的弊端。私立学校和公立学校的好坏是一个实证性问题，可以使用社会科学中的调查研究方法来辨别真伪。但是，这并不意味着只要这个问题是个科学问题并有可能获得解决，就是一个非常简单的问题。所有鼓吹私立学校优越性的人都潜在地意识到这一点，因为他们在维护自己的观点时，常常引用这样一个经验性的事实：私立学校学生的成绩要好过公立学校。尽管这个事实无可辩驳——各种研究中有大量一致的教育统计数据，但问题在于，用这些学生的成绩数据就推出结论，即私立学校的教育本身导致了较高的分数，这么做是否合适？

考试成绩是许多不同变量的函数，这些变量彼此之间又是相关的。为了评估公立学校和私立学校的好坏，我们需要进行更为复杂的统计，而不仅仅是学校类型和学业成就之间的相关。例如，学业成就和家庭背景中许多不同指标都有关系，如父母的教育程度、父母的职业、社会经济地位、家中藏书的数量以及其他一些因素。这些特征都与是否把孩子送到私立学校有关系。因此，家庭背景是一个潜在的第三变量，可能会影响到学业成就和学校类型之间的关系。简而言之，学业成就可能和学校质量没有任何关系，而结果可能是，家境优越的孩子学习更好，更有可能进入私立学校。

幸运的是，还有许多复杂的相关统计方法，例如多元回归、偏相关、路径分析（统计学的发展部分要归功于心理学家），这些复杂的统计方法能够去除其他变量的影响、提出公因子或定义协变量之后，重新计算两个变量之间的相关。运用这些更为复杂的相关技术，研究者对大量有关高中生的教育统计进行了分析，结果发现，当反映学生家庭背景和一般智力能力的变量被排除后，学业成就和学校类型之间几乎就没有一点关系了。其他研究者也确认了他们的研究结果（Berliner & Biddle, 1995; Carnoy, Jacobsen, Mishel, & Rothstein, 2005; Hendrie, 2005）。

因此，很明显，鼓吹私立学校能够提高教育成就，就跟讨论节制生育需要用“烤箱”一样没什么分别。学业成就和私立学校相关，不是因为任何直接的因果机制，而是因为私立学校中学生的家庭背景和一般认知水平与那些进入公立学校的学生相比是不一样的。

这些较为复杂的相关统计方法能够排除第三变量的影响，但并不总是会削弱原有相关的强度。有时候，在排除第三变量之后，两个变

量之间的原有相关仍然存在，这个结果本身就能提供一些信息。这样的结果说明，原有相关并不是由第三变量所导致的虚假相关。当然，并不排除其他变量也会导致虚假相关。

另一个例子，有研究发现，高中生能否升入大学和这个学生的家庭社会经济地位有关。这是一个重要发现，足以动摇我们这个社会的核心价值——实现目标靠的是个人能力。它表明，一个人的成功取决于这个人的经济地位。但是在下这个结论之前，我们必须首先考虑一下其他假设。这就是升入大学和社会经济地位之间的相关可能是一种假象。其中一个非常明显的第三变量就是学业能力，它可能与升入大学和社会经济地位都有关系，如果这个变量被排除出去，这两个变量之间的相关就会消失。然而，从研究者运用正确的方法计算出的数据中发现，在学业能力被排除后，升入大学和社会经济水平的相关仍然显著（Baker & Velez, 1996; Long, 2007）。因此，高收入阶层的孩子更容易进入大学不能完全归因于学业能力的不同。当然，这个发现不能排除这种可能性：其他一些变量导致了升入大学和社会经济水平之间的相关，但是能够用这样一种再分析来排除学业能力对两者相关的影响，这本身就在理论及实践方面具有很重大的意义。

还有一个使用偏相关技术的例子。安德森等（Anderson & Anderson, 1996）描述了他们是如何来检验关于暴力的地区差异理论的，他们通过检验一系列不同的理论看其是否能够对所收集的数据作出解释。他们采用偏相关技术来进行此项研究。曾有研究表明美国南部地区的暴力犯罪高于北部地区，他们检验了“热假设”——令人不适的高温增强了侵犯性动机和攻击性行为（p. 740）。他们发现城市平均气温和暴力犯罪率之间存在相关，这并不令人奇怪。但是从统计上控制一些变量，如失业率、个人平均收入、贫困率、教育程度、人口规模及其他一些变量之后，气温和暴力犯罪之间的相关仍然显著。这就使得“热假设”理论的可信度大大提高了（又见Larrick, Timmerman, Carton, & Abrevaya, 2011）。

方向性问题

如果能够采用某种方式操纵变量，并能够因此作出科学的因果推断，就没有理由仅凭相关证据作出因果推论。而让人苦恼的是，当涉及心理学主题时，仅根据相关就得出结论的做法却是普遍存在的现象，在心理学知识对解决社会现实问题愈发重要的今天，这种倾向所造成的损失也与日俱增。在教育心理学界，一个广为人知的例子很好地诠释了这一点。

自从一百年前关于阅读的科学研究开始以来，研究者们就知道，眼动模式和阅读能力之间存在着相关。阅读能力差的人，其眼动轨迹是不规则的，表现为更多的回扫（从右向左的运动），在每一行上的注视时间（停顿）更长。基于这种相关，一些教育工作者假设，眼球运动技能的缺失是造成阅读问题的原因，因此许多“眼球运动训练计划”在小学生中展开和实施。在查明这一相关是否真的说明“不规则的眼球运动会导致低下的阅读能力”之前，这些训练计划已经开展了很长时间。

现在已经清楚了，眼球运动与阅读能力的相关反映了一种与之前所想象的完全相反的因果关系。不规则的眼动并不会导致阅读障碍，相反，是缓慢的单词识别和理解困难导致了不规则的眼动。当教会儿童有效地识别单词和更好地理解文字后，他们的眼动轨迹就变得平顺了。训练儿童的眼球运动和提高其阅读能力是没有关系的。

最近十几年，研究者们已经明确指出，文字解码和语音加工方面的语言问题是阅读障碍存在的根源（Snowling & Hulme, 2005; Stanovich, 2000; Wagner & Kantor, 2010），而几乎没有眼动模式导致阅读障碍的案例。但是，如果到大部分中等规模以上的学校的储藏室里仔细翻一翻，都能找到布满灰尘的眼球运动训练仪器，这表明数以千计的买设备的钱被浪费了，这就是把相关视为因果证据的后果。

考虑另一个类似的例子。在教育和社会服务领域里有一个非常流行的观点：学业成就问题、药物滥用、青少年怀孕以及其他一些问题行为都是低自尊造成的。这一说法假定，此因果关系的方向很明显：低自尊导致行为问题，高自尊带来高的学业成就和其他领域的成绩。

这种方向性因果关系假设为许多提高自尊的教育计划提供了动力，这个问题和眼球运动那个例子是一样的：仅仅因为存在相关就推出一个方向性的因果假设。事实证明，就算真的存在因果关系，自尊和学业成就之间的关系更可能呈相反的方向：高学业成就（包括生活中其他方面）导致了高自尊，而不是反过来（Baumeister et al., 2003, 2005; Krueger et al., 2008）。

在心理学研究中，因果关系方向的确定也是一个普遍的问题。例如，心理学家乔纳森·海特（Jonathan Haidt, 2006）讨论了一项研究，该研究发现利他和幸福有关。例如，有研究表明，做志愿活动的人要比不做志愿活动的人更幸福。当然，确定没有第三个变量能解释利他和快乐之间关系是必要的。一旦第三变量被消除，决定关系的方向就非常重要了。是幸福让人们更利他，还是利他行为让人们幸福（“赠与比接受更值得赞美”）？当依照第6章所述的真实验的逻辑，进行了适当的控制研究，发现在两个方向上都有关系：幸福让人们更多地利他，并且利他行为也会让人更幸福。

到目前为止，我们的讨论主要围绕变量间相关所涉及的两种误区：其中一种叫做方向性问题，已经通过眼球运动和自尊的例子进行了阐述。当变量A和变量B之间存在相关时，在断定A的变化引起B的改变之前，我们必须清楚因果关系的方向可能是相反的，即从B到A；第二种是有关第三变量的问题，此问题已经通过糙皮病的例子（以及烤箱—节育和私立学校—学业成就的例子）加以论述。两个变量之间的相关并不能预示任何方向上的因果，因为当这两个变量都和第三变量相关时，该相关还是会出现。

选择性偏差

在一些情境下，虚假相关很容易出现。这也正是选择性偏差非常容易出现的原因。“选择性偏差”这个术语指的是特定主体和环境变量之间的关系，当不同生理、行为、心理特点的人们选择不同类型的环境时，就有可能出现选择性偏差。选择性偏差造成环境特征和行为 —生物特征之间的虚假相关。

让我们通过一个例子来了解选择偏差是如何产生虚假相关的。请快速说出一个州名，在这个州里，由呼吸系统疾病导致的死亡率高于平均水平。当然，答案之一是亚利桑那州。什么？等等！难道亚利桑那州没有清洁的空气吗？难道洛杉矶的烟雾弥漫得如此之远？难道凤凰城的郊区环境已经变得那么差了吗？不是，肯定不是！让我停下来想一想。可能亚利桑那州的确有清洁的空气，可能患有呼吸疾病的人都愿意搬到这里，然后他们死在了这里。这样就对了。如果我们不够认真，就会出现上面所说的那种情形：我们可能会受到误导，认为是亚利桑那州的空气害死了这些人 [1] 。

但是，选择性偏差并不总是那么容易辨别。尤其是当我们事先就期望看到因果联系时，这种偏差经常会被忽略，就像在“自尊”的例子中那样。充满诱惑的相关性证据加上固有的偏见，就能够欺骗最聪明的头脑。下面让我们看一些事例。

从关于“美国教育质量”的全国性讨论中，可以很容易地看到选择性因素的重要性，这场讨论已经在美国全国范围内持续了近二十年。在这场辩论中，公众被各种教育统计数据所淹没，但研究者却没有提醒公众，警告他们避免从相关数据去推论因果关系，因为相关数据内含有大量具有误导性的选择性偏差。纵观这场辩论，许多怀有政治目的的人试图不断地提出证据，用以说明教育质量和教师的薪资水平、班级规模是没有关系的，尽管已有许多研究表明这二者都非常重要（Ehrenberg, Brewer, Gamoran, & Williams, 2001）。他们所提到的证据当中，有一个是50个州的SAT（学术能力评估测试）成绩。这个测试的参加者是有意升入大学的高中生，学生的平均测试分数确实表明，学生成绩和教师薪资水平、教育的支出是没有关系的。即使有关系，其趋势看起来也与期望的方向相反。在许多州，教师薪资水平很

高，但是SAT的测试成绩很低，有些州教师的薪资水平在全国工资水平排行垫底，而学生的SAT测试成绩却很高。对这组数据的仔细审视给我们上了另外一课：选择偏差导致虚假相关是多么容易。

在进一步的检验中，我们看到，密西西比州学生在SAT考试中的得分高于加利福尼亚州学生（Grissmer, 2000; Powell & Steelman, 1996），而且差异是非常显著的，密西西比州比加州的平均分要高出100分。而密西西比州的教师薪资水平在全国是最低的，这无疑会让那些鼓吹削减教师工资的人们弹冠相庆。但是，请等一下！密西西比州的学校真的好于加利福尼亚州？前者的教育水平真的高于后者？当然不是。几乎任何一个客观的指标都显示，加利福尼亚州的学校更好。但是如果这是真的，那么SAT的成绩又该如何解释？

这个问题的答案要用选择性偏差来解释。SAT和学校通常选择的那些标准化考试不同，在标准化考试中，所有学生一律都要参加。但SAT并不是所有的高中生都参加的，因而存在选择性偏差。只有那些希望进入大学的学生参加这个考试。这个因素就能够解释州与州之间的平均分为何存在差异，同时解释了为什么一些州有最好的教育体制，在SAT考试中的平均分却很低。

选择性因素在两个方面操纵了SAT的得分：首先，一些州立大学需要ACT（美国大学考试）的成绩，而不是SAT分数。所以在这些州中，只有那些打算去州外读大学的学生才会参加SAT考试。比起那些平均水平的学生，这些学生中的大部分最有可能拥有更好的家庭条件或者更高的学术才能。这种情况也发生在密西西比州和加利福尼亚州的考试中。密西西比州仅有4%的高中生参加SAT，然而加利福尼亚州却高达47%（Powell & Steelman,1996）。

第二个选择性因素则更加微妙。在那些教育质量高的州里，许多学生在高中毕业后，更倾向于继续接受教育。在这些州，参加SAT考试的学生比例高，这其中也包括一些学习成绩较差的学生。而在那些有着高辍学率、低教育质量的州中，想继续接受大学教育的学生比例很低。在这些州中，最终参加SAT考试的学生代表的是这些州中学习成绩比较好的那些人。因此，他们的平均成绩自然要高于那些大部分人都参加升学考试的州。

关于SAT分数的这个例子也为我们提供了一个反面教材，那就是：公众如果缺乏本书所教授的简单方法论和统计思维技能，想纠正那些误导性的数据会何等地困难。印第安纳州的教授布赖恩·鲍威尔（Brian Powell, 1993）分析了由政治专栏作家乔治·威尔（George Will）在1993年所写的一篇专栏文章，威尔反对公共教育支出，因为

在SAT测试中取得高分的州，并没有高的教育支出。鲍威尔指出，威尔挑出的那些SAT分数特别高的州——爱荷华州、北达科他州、南达科他州、犹他州和明尼苏达州——参加SAT考试的学生比率分别为5%、6%、7%、4%和10%，然而在美国参加SAT考试的总比率是40%以上。原因就是，在以上这些州中，要想进入公立学校，必须参加ACT考试，只有那些计划去州外有名望的私立学校读书的学生才参加SAT考试（Powell, 1993, p. 352）。与之相反，在威尔列举的新泽西州，SAT分数很低，教育支出却很高，其中有76%的高中生参加了这个考试。显然，相比新泽西州，在南、北达科他州参加SAT考试的学生配称得上是一支“精锐之师”。

当存在选择效应的时候，妄下结论可能会导致我们做出错误的现实选择。许多更年期之后的女性曾经被怂恿去尝试激素替代疗法（HRT），因为有报道说这一疗法可以降低心脏疾病的概率。不过能够表明这一点的早期研究仅仅是把一组选择进行HRT（自我选择进行治疗）和一组没有进行HTR治疗的女性进行对比。然而，真正进行实验（进行随机分配，见第6章）时发现，实际上HRT根本就不能降低心脏病的概率（Bluming & Tavris, 2009；Seethaler, 2009）。包含了特定样本的早期研究之所以说明了HRT确实有效，是因为选择HRT的女性比不选择HRT的女性本来就更加积极进行锻炼，不那么肥胖，更不可能抽烟。

来自临床心理学的例子可以表明，选择性偏差问题是多么具有欺骗性和违背常理。研究数据有时会显示，接受心理治疗的人在各种成瘾症——如肥胖、吸毒、吸烟的治愈率方面，要低于那些没有接受过心理治疗的人（Rzewnicki & Forgays, 1987; Schachter, 1982）。你想知道原因吗？原因并不是因为心理疗法使得成瘾的行为更加难以改变，而是因为那些寻求心理治疗的人的成瘾问题更加复杂和棘手，而且很少能够自愈。简而言之，“困难的问题”比“容易的问题”寻求心理治疗的几率更高。

维纳（Wainer, 1999）给我们讲了一个二战期间的故事，这个故事提醒我们选择性偏差违背常理的一面。他提到一位飞机分析师，这个分析师一直试图通过分析飞机被子弹击中的弹孔分布，来确定飞机上的哪个部位是应该放置加固防弹层的位置。他最后的决定是：把加固防弹层安放在返航机上没有弹孔的地方。他的理由是，子弹袭击飞机各个部位的几率是均等的，所以，如果一架飞机能返回，就表示这架飞机被子弹击中的地方必定不会对飞机造成致命损伤。看来那些没有弹孔的地方都是要害，因为该部位如果被击中，飞机可能就无法返航。因此，加固防弹层应该安装在返航机没有被击中的部位！

使用选择效应很容易“诱骗”人们作出因果推论。让我们看看这个：共和党比民主党更加享受性爱。这绝对是事实。统计资料显示，共和党选民平均来说比民主党选民对他们的性生活更加满意（Blastland & Dilnot, 2009）。究竟是什么能让拥护共和主义的人更性感呢？

没错，那不对。党派不会改变一个人的性生活。那么，怎么对数据进行解释呢？两个方面：第一，投票支持共和党男性要多于女性；第二，调查显示男性比女性对性生活更加满意。共和主义没有改变任何人的性生活，而仅仅是因为有更高满意度的人群（男性）更加倾向于为共和党投票。

像“更加性福的共和党人”这样的例子告诉我们，当选择效应起作用的时候要千万小心。经济学家史蒂文·兰兹伯格（Steven Lansburg, 2007）向我们说明了与使用技术相关的生产率的数据大部分可能存在对因果关系的过分解释，而实际上控制选择效应之后它们只是相关关系。在公司里，往往是最具生产力的员工获得最先进的技术。因此，当计算相关时，生产率和技术使用相关。但是，不是技术改善了这些员工的绩效，是在接受到先进技术指导前，他们本身已经更具生产力了。

涉及选择效应的一个重要的现实健康问题就是饮酒对健康的影响。大量的研究发现，中度饮酒者比频繁饮酒者和不饮酒者更加健康（Rabin, 2009）。我们知道了选择效应，不论你我都不会去劝诱戒酒者如果他们少喝点酒就会改善健康状况。这是因为自主去选择喝酒的人会控制他们的饮酒量。就像拉宾（2009）解释的，已经发现中度饮酒者对他们做的任何事情都是适度的。他们适度锻炼、适度进食。他们倾向于做许多正确的事情。所以我们当然不知道导致了积极的健康后果的是适度饮酒本身，还是其他健康的适度特征（锻炼水平、节食，等等）。由于选择效应，我们不能说适度饮酒本身就是原因。

总之，这一章提供给读者的规则很简单：提防选择性偏差的发生；当只有相关时，应避免因果推论。不可否认，复杂的相关数据里确实存在着有限的因果关系。同样不可否认的是，相关的证据有助于证明假设的聚合效度（见第8章）。然而对于心理学知识的消费者来说，宁可站在怀疑的角度，也不要被那些错误地暗示了因果关系的相关所蒙蔽。

注释

[1] 亚利桑那州在美国西部，以地广人稀、气候干燥、空气清洁著称。——译者注

小结

本章的主旨是想传达这样一个理念：两个变量之间仅仅存在相关，并不能保证一个变量的变化就会导致另一个的变化，关键就在于相关并不意味着因果关系；在第三变量问题里，两个变量之间的相关并不意味着它们之间存在直接因果路径，因为相关的产生可能是由于这两个变量或许都与未被测量的第三变量有关。事实上，如果潜在的第三变量也经过了测量，就可以用相关统计，如偏相关（第8章将会讨论）来评估第三变量是否决定了这种关系；让相关统计的解释变得困难的另外一个原因就是方向性问题。实际上，如果两个变量有直接的因果关系，因果关系的方向是不能根据相关来判断的。

在行为科学中，选择性偏差是造成诸多虚假相关的罪魁祸首。事实上，人们在一定程度上选择他们的环境，并人为创造了行为特性和环境变量之间的相关。正如戈德伯格的例子所阐述的那样（在接下来的两章中，我们将会进一步讨论），确保选择性偏差不会捣乱的唯一方法，就是在操纵所有变量的情况下进行真正的实验。

Chapter 6
让一切置于控制之下：聪明汉斯的故事

这一章开始前，咱们先来做一个小测验。噢，别担心，不是考你前几章所学的内容。问题其实很简单，是有关现实世界中常见的物体运动方面的知识，问题只有三个。

首先，你需要一张纸，想象如下场景：一个人拿着一根细绳在他的头顶上绕圈，绳子的另一端系着一个球。画一个圆来代表从上方俯瞰这个球的运动轨迹。在这个圈的一处画一个点，然后用一条线把这个点和此圆的圆心连接起来。这条线就代表那根细绳，那个点就代表特定时刻的球。想象在某一旋转瞬间，细绳断了。你的第一项任务是用笔画出这个球飞出后的运行轨迹。

第二个问题，假设你是一个轰炸机的飞行员，现在正以每小时500英里的速度在20000英尺（约6096米）的高空飞向目标。为了简单起见，假设没有空气阻力，问题是：什么地方是投掷炸弹的最佳位置，是在到达目标地点之前，还是目标的正上方，或者是在你经过目标之后？无论你选择的是目标之前、目标正上方，还是飞越了目标之后，都请你指出投放点与目标的具体距离。

最后，想象你正拿着一把来复枪从肩膀高度处开火。假设没有空气阻力，且步枪与地面是平行的。如果子弹从与枪相同的高度落地需要1.5秒钟的时间，那么假设你现在由枪管中射出一发子弹，初速度是每秒2000英尺（约609.6米），那么子弹落地需要多长时间？

答案——对了，还有答案这回事儿。答案会在本章的后面揭晓。

但在此之前，为了便于理解掌握这些运动方面的知识与心理学有什么关系，我们需要先深入地探讨实验逻辑的本质，这些实验逻辑经常被科学家们所使用。在本章，我们将要讨论实验控制和操纵的一些原理。

斯诺与霍乱

在前一章我们讲到，约瑟夫·戈德伯格对糙皮病的研究在一定程度上是受“糙皮病是不会传染的”这种预感的指引。但是比戈德伯格早70年，约翰·斯诺（John Snow）在对霍乱起因的研究过程中则将病因放在相反的猜想上，但同样获得了成功（Johnson, 2007; Tufte, 1977）。早在19世纪50年代的伦敦，人们对不断暴发的霍乱提出了许多理论，并且彼此争论不休。很多医生认为霍乱病人呼出的气体会将此疾病传染给别人，此理论被称为“秽气理论”。但斯诺却提出，该疾病是通过被病人排泄物污染的供水系统传播出去的。

斯诺开始着手验证他的理论。幸运的是，当时伦敦有许多不同的供水源，每个供水源给不同的地区供水，所以不同供水系统受感染的程度不同，霍乱的发生率应该因供水源受污染程度的不同而存在差别。但是斯诺发现，这种比较会出现严重的选择性偏差（请回想一下第5章的讨论）。在伦敦，不同地区的贫富差距非常大，因此，供水系统和各地区患病率之间的任何相关都会受到其他能够影响健康的、与该地区的经济发展水平相关的变量的影响，如饮食、压力、工作危机或生活质量。简而言之，获得虚假相关的可能性很大，这和第5章所讨论的糙皮病和污水的关系类似。但是斯诺非常机敏地注意到了一种已经出现过的特殊条件，并利用这一点解决了问题。

在伦敦的一个市区，碰巧有两家自来水公司对同一个社区供水，但从供水布局上来说是杂乱无章、毫无规划的。在某条街道上，一部分住宅是由其中一家自来水公司供水，一部分是由另外一家自来水公司负责供水，这种情况发生的原因是由于最初两家公司存在竞争。甚至有这样的情况，一栋房子由一家公司供水，而与它毗邻的房子却是由另一家公司供水。因此斯诺找到了几个由两家公司分别供水的家庭，并且这些家庭的社会经济地位基本相同，或至少是非常接近的。如果两家自来水公司都受到污染，那么这种选择仍旧是没有任何意义的，因为这样斯诺就不能发现水污染与霍乱的发病率有什么关系了。所幸的是，这种情况并没有发生，这两家公司的水并未同时受到污染。

在一波霍乱流行过后，兰姆博斯（Lambeth）公司为了避免水污

染，将公司迁到泰晤士河的上游，而南沃克—沃克斯霍尔（Southwark & Vauxhall）公司却仍然固守在下游。因此，兰姆博斯公司的水系统受污染的可能性比南沃克—沃克斯霍尔公司要小得多。斯诺通过化学检验也证明了这一点。剩下的工作就是统计由两家不同公司供水的家庭的霍乱发病率：兰姆博斯公司供水的每10000个家庭里有37人死亡，南沃克—沃克斯霍尔公司供水的每10000个家庭里有315人死亡。

在这一章我们要讨论的是，斯诺和戈德伯格的故事是如何体现科学思维的逻辑性的。如果不能理解这种逻辑性，科学家们的所作所为看上去就会显得很神秘、怪异或是荒唐透顶。

比较、控制和操纵

尽管市面上关于科学方法论的书已经汗牛充栋，但是对于从未做过实验的外行人士来说，这些书可能都如同浮云一般，因为外行人只想知道一个大概，并不想搞清楚实验设计的所有复杂细节。科学思维最重要的特点很容易掌握，那就是科学思维所基于的理念是比较、控制和操纵。要想获得对一个现象更加深入的了解，科学家就要比较世界上存在的各种情况。没有这种比较，我们所观察到的都是一些孤立的事件，并且对这些孤立的观察结果也解释不清，就像我们第4章所讨论的见证叙述和个案研究一样。

科学家通过比较在不同条件（但是有控制的）下得到的结果，可以排除一些错误的解释，并证实正确的解释。实验设计的基本目的是分离变量。当成功分离出一个变量，实验的结果就能排除大量之前提出作为解释的其他理论。科学家们通过两种方法尽可能地排除不正确的理论：要么是在实验条件下直接进行控制；要么在自然情境下进行观察，以便比较各种可能的解释。

后一种情形在霍乱这个例子中得到了很好的诠释。斯诺并不是简单地随意选择两家自来水公司，他清楚自来水公司可能给不同地区供水，并且这些地区的社会经济水平会有很大差异，这种社会经济水平的差异很有可能会影响人们的健康水平。仅仅观察不同地区霍乱的发病率，难以避免“同时存在许多不同解释”的问题。斯诺清楚地知道，科学的不断发展需要尽量减少对同一个问题的各种不同解释（请回想一下第2章所讨论的可证伪性），因此他不断寻找并且最终找到一种比较方式，此方式可以排除一大堆解释，这类解释都是与健康有关的社会经济地位方面的因素。

斯诺幸运地找到了一种自然情境，这种情境使得他能够排除其他的可能性。这种在自然情况下产生的“比较”条件并不多见。让科学家坐在那里等待这类情况发生是十分荒谬的。事实上正相反，很多科学家都试图以一种区分各种不同假设的方式来重构世界。为实现这一目的，他们必须操纵被认为是诱因的变量（在斯诺的实验里是被污染的供水系统），然后在保持其他所有相关变量不变的情况下，观察是否会有不同的结果（霍乱的发病率）。被操纵的变量称为自变量，随着

自变量变化而变化的变量称为因变量。

因此，一个好的实验设计应该是这样的：科学家能够操纵他感兴趣的变量，并对其他可能影响实验的无关变量进行控制。需要注意的是，斯诺并没有这么做。他不可能操纵供水系统的污染程度，但是他找到了这样一种条件，即供水系统受污染的程度是不同的，并且与社会经济水平有关的其他变量侥幸得到了控制。可是这种自然发生的情境不仅很少见，而且也不如直接的实验操纵那么具有说服力。

约瑟夫·戈德伯格就是直接操纵变量，他假设这个变量就是引起某种特别现象的原因。戈德伯格不仅对与糙皮病相关的变量进行观察和记录，他还在一系列研究中直接操纵了其他两个变量。回想一下，他安排了低蛋白饮食的囚犯组来诱发糙皮病，同时安排吞食糙皮病患者排泄物的志愿者，其中还包括他妻子和他自己。因此，戈德伯格不仅观察了自然发生的情境，还创设了特殊条件组，从而排除一系列其他可能性并获得实验结果，这种推论要比斯诺的方法更具说服力。这也正是为什么科学家要试图操纵一个变量并保持其他所有的变量不变的原因：为了排除其他的可能性。

随机分配与操纵共同定义了真实验

我们这里并不是说斯诺的方法毫无可取之处。但科学家们的确愿意更为直接地操纵实验变量，因为直接操纵变量能够产生更具说服力的推论。细想斯诺的两组被试：一组由兰姆博斯公司供水，另一组由南沃克—沃克斯霍尔公司供水。由于处在同一个地区，可能保证了两组被试的社会地位几乎相同。

但是类似斯诺这类实验设计的缺陷是，它是由被试决定自己属于哪一个组的。因为他们早在几年前已与两家自来水公司签订了供水合同。我们还必须考虑为什么一些人与这家公司签约，而另外一些人与那家公司签约。是不是一家公司比另外一家公司的口碑好？是由于这家的价钱比较便宜，还是广告说这家的水有很好的药用价值？我们不得而知。关键的问题是，这些人选择其中一家公司是不是因为该公司做广告说他们的产品质量优于另外一家，特别是对人的健康有益处？而或许这些因素才是低发病率的真正原因。这是有可能的。

类似斯诺这样的实验设计就无法排除那些更为微妙的虚假相关，这类虚假相关不像其他与社会经济地位有关的相关那样容易被看出来。这就是科学家倾向于直接操纵他们感兴趣的变量的原因。当操纵

变量与一种叫做随机分配的程序（在随机分配中被试不能决定自己进入哪种实验条件，而是被随机分配到某一个实验组）相结合时，科学家们就能够排除那些可以归因为被试本身特征的解释了。随机分配确保被试在对比实验条件下的所有变量基本保持一致，随着样本数量的增加，它还能平衡掉一些偶然因素。这是因为被试的分配是由不带偏见的随机方法实施的，而不是由某个人的选择决定的。请注意这里的随机分配与随机样本不是一回事，这两者的区别我们将会在第7章进行讨论。

随机分配是一种将被试分配到实验组和控制组的方法，以保证每个被试有同样的几率被分到其中一个组。掷硬币就是一种决定某一被试分到哪一组的手段。实际实验中往往采用电脑生成的随机数字表。通过使用随机分配，研究者在研究之前就试图平衡两组的所有行为变量和生理变量，甚至是那些研究者没有进行专门测量或考虑到的变量。

随机分配的效果如何，取决于实验中被试的数量。也许你会认为被试越多越好，也就是说，分配到实验组和控制组的被试的数量越多，两组间除了自变量以外的其他所有变量就越接近。但幸运的是，对于研究者来说，其实每组只需要一个相当少的人数（例如20～25人），随机分配就可以起到很好的效果。

使用随机分配能有效避免由于分组方式所导致的系统误差。这两组被试在所有变量上均得到匹配，但即使存在一定程度的不匹配，随机分配也消除了实验组或控制组之间的偏差。如果我们了解一下“重复”这个概念，对于随机分配如何去除系统误差这个问题就比较好理解了，所谓的重复是指在各种环境下重复一个实验，看还能否得到同样的实验结果。

设想一下，一个发展心理学家想要做一个关于早期丰富体验对学前儿童的影响的实验，在日托期间，随机分配到实验组的儿童每天接触心理学家设计的大量丰富活动，随机分配到控制组的儿童在同样的时间里只是参加一些比较传统的游戏活动。因变量是儿童上学一年后的期末成绩，通过成绩考察实验组儿童的表现是否优于控制组儿童。

像这样的实验就会用到随机分配，以确保两组在实验之初，所有能够影响因变量的无关变量都基本保持一致。这些无关变量有时被称为干扰变量。这个实验中的干扰变量可能会是儿童的智力测验成绩和他们的家庭环境。随机分配将会在大体上使两组间在这些变量上保持平衡。但也有例外，尤其当被试人数很少时，两组仍然有可能存在差异。例如，如果随机分配之后，实验组儿童的智力测验的成绩是

105.6, 控制组的是101.9（尽管恰当地使用了随机分配，这种差异还是有可能发生），我们就会担心实验组的学业成就的任何变化缘于该组儿童的智力测验成绩高，而不是由于他们经受了丰富的体验。这里就能看出重复验证的重要性了。后续研究进行随机分配之后，两组仍然可能存在智商差异，但是随机分配程序避免了系统误差，这就能够保证这种差异不会总是出现在实验组。事实上，无系统误差这一点所确保的是，在一定数量的类似研究中，智商差异出现在实验组和出现在控制组的概率是相等的。在第8章中，我们将会讨论如何使用这种多重的实验来提高结论的聚合效度。

因此，随机分配程序有两个优点。一个是在任何实验中，样本的数量越大，随机分配越能平衡两组所有其他的无关变量。而即使在一些匹配得不是特别好的实验里，由于随机分配克服了系统误差，仍然可以让我们得出令人信服的结论——只要研究可以被重复。这是因为，经过一系列这样的实验，两组间混淆变量造成的差异就会被平衡。

控制组的重要性

科学研究中不乏由于缺乏真实验的完全控制而得出错误结论的例子。罗斯和尼斯贝特（Ross & Nisbett, 1991）提到一种多年前非常流行的治疗肝硬化的疗法——门腔静脉分流术的医疗发现。1966年人们开始对此疗法进行大量研究，并且发现了一种引起人们兴趣的现象。在96.9%的不包含控制组的研究中，医生判断这种治疗方法的效果至少在中等程度以上。在有控制组但没有使用随机分配的研究中（因此不属于真实验设计），86.7%的研究显示同样的结论。但是，在有随机分配的控制组的研究中，只有25%的研究显示同样的结论。因此在今天，这种特殊治疗方法被认为是无效的，但在当时，由于没有进行完全的实验控制，治疗效果被夸大了。罗斯和尼斯贝特（1991）指出："没有使用较为正式的实验程序所获得的积极效果，要么是'安慰剂效应'的产物，要么是由于没有使用随机分配而产生的偏差。"（p.207）罗斯和尼斯贝特还继续探讨了"当没有使用随机分配的时候，选择性偏差是如何产生虚假相关的"这一问题。例如，如果一些病人被选作某种治疗方法的研究被试，他们可能会努力做一名好的参与者，或者他们拥有家庭的支持、积极的态度或者他们的家人对其病情更为关心，这些都可能影响实验组与控制组的差别，而这与治疗方法的效果没有任何关系。

在下结论之前，必须获得"比较信息"，这种思维倾向并不是与生

俱来的，这就是为什么所有科学研究都要经过训练。这些训练包括强调控制组的重要性的研究方法课程。控制组和实验组很像，只不过缺少一种重要因素的影响。

控制组的这种“非鲜明性”很难让人发现它的重要性。心理学家们做了大量的研究来说明人们为什么忽视重要的比较（控制组）信息。例如，在一个研究范式中（Stanovich, 2010），我们给被试呈现一个2×2的实验数据矩阵，如表6-1所示。

表6-1

	好转	没有好转
接受治疗	200	75
未接受治疗	50	15

表6-1中的数字代表每种情况的人数。具体来说，200人在接受了治疗后表现出病情好转，75人接受治疗但没有任何好转，50人没有接受治疗但仍有好转，15人没有接受治疗也没有任何好转。研究者让看过这一矩阵的被试指出治疗是否有效，很多被试认为测试中的治疗方法是有效的，相当多的被试甚至认为治疗是很有效的。这是因为他们首先关注的是200人接受了治疗且好转的那一组。其次，他们关注这样一个事实，即接受治疗且好转的人数（200）要远远多于没有好转的人数（75）。

事实上，这个实验所检测的疗法是完全无效的。为了理解为什么这个疗法是无效的，有必要关注一下表示没有接受治疗的控制组（没有接受特殊疗法的组）的两格数据。我们可以看出，控制组的65人中有50个人，即76.9%的人即使没有接受特殊治疗还是有所好转。这与275人中200人（72.7%）接受治疗且有所好转形成了对比。因此，控制组中病情好转者的比例实际上更大，这说明这种疗法是完全没有效果的。只关注实验组的结果而忽视控制组的结果，会诱使许多人认为这种疗法有效。简而言之，它很容易让人们忽略这一事实，即当我们对治疗效果进行解释时，控制组的结果是背景信息中极为关键的一环。

不幸的是，我们的媒体经常干的事情，就是将人们的注意力从比较性信息的必要性上移开。心理学教授彼得·格雷（Peter Gray, 2008）谈到在《时代》（Times）杂志上一篇题为《离婚的持久性伤害》的文章，文章列举了很多历史上的案例，报道了许多父母离婚的人。当然，在缺乏非离异家庭个体的控制组时，我们不能从这里得出任何结论。我们怎么知道离异家庭的个体更可能表现出这些消极的结果呢？只有一个匹配的控制组才能回答这个问题。

除了这类例子，社会学和不同的应用学科在评估证据时，也开始越来越重视比较性信息的必要性。这里列举医学领域近期还在进行中的一个研究进展（Gawande, 2010; Redberg, 2011）。神经学家罗伯特·伯顿（Burton, 2008）描述了医学采取的路径——从对人造成伤害的直觉知识，到建立在比较研究所获得的有用知识上的治疗方法。“多年来，我常常感到惊讶，为什么许多聪明、训练有素的医生会进行一些不必要的外科手术，而且这些手术未经验证，又有危险。在医学实践的核心里存在一个巨大的矛盾：我们从经验中学习，但是如果没有经过足够的实验，我们就无法知道我们对一个特定治疗结果的解释是否正确……但是，当一名好医生就需要坚持最佳的医学证据，即使它和你的个人经验相矛盾。我们既需要区分直觉和可检验的知识，也要区分预感和经过实证检验的证据。”（pp.160—161）

其他实践领域中的直觉性“预感”也越来越多地被置于控制对比研究中，以进行检验。例如，信用卡公司经常寄出信件，提供可选择的条款，判断哪个条款对客户最有吸引力（Ayres, 2007）。例如，一组随机分配的家庭会收到一个利率、年费和奖励计划的组合。另一组随机分配的家庭会收到另一个不同的利率、年费和奖励计划的组合。如果两组的接受率上有差异，那么公司就会发现哪个组合更好（从吸引更多客户的角度）。重点在于，信用卡公司无法获悉其现行的条款是否“起作用”（例如，是否吸引了尽可能多的客户），除非他们进行一些实验，将可选择的条款进行比较。

不仅在商业方面，政府部门也开始使用控制实验来探寻如何进行政策优化。美国住房和城市发展部进行了一个实验，叫做“向机遇迁居实验”（Ayres, 2007）。对一组随机分配的低收入家庭给予住房代金券（可以用在任何地方）；给予另一组随机分配的低收入家庭的代金券只能用在低贫困（例如中产阶级）地区。这样做的目的是了解当低收入家庭的周围不是其他低收入家庭时，在结果变量（教育成果、犯罪行为、健康状况，等等）上是否有差异。这种类型的研究被称为“现场实验”——变量的操控是在非实验室条件下进行的。另一个政府赞助的现场实验是墨西哥教育、健康和营养改善项目（Ayres, 2007）。这个项目包括有条件地将钱转移给贫困家庭。当母亲接受产前检查时，她们就可以得到现金。她们的孩子入学并通过了营养检查，也可以得到现金。

政府在506个村庄进行了现场实验，以验证这个项目的功效。半数的村庄参与了这一计划，而另一半没有。这使得政府能够检验该项目的成本效率。两年后，对这些村庄的成果进行检查，例如教育成就、营养和健康水平。如果没有控制组，政府就无从得知在没有该计

划的情况下，教育和健康会是什么样的水平。

国际援助组织也致力于有操纵变量的研究（真实验）来找出“什么起了作用”（Banerjee & Duflo, 2009）。作家尼古拉斯·克里斯托夫（Nicholas Kristoff, 2009）论述了援助组织的问题，他们经常进行自我评估并最终声称他们做的所有事情都起了作用，而这是不现实的。这种路线方法意味着钱将会花错地方。为了更有效地利用援助资金——即拯救更多的生命——判断哪个项目比其他项目更有效是非常重要的。克里斯托夫描述了麻省理工学院的扶贫行动实验室是怎样设计研究的，为了找到哪个项目是最有效的，研究至少是合适的真实验，随机在一些地方进行援助计划而在其他地方不进行。

有时公众很难理解，他们想要的无非是有效地利用纳税人的钱帮助大多数人，为什么必须要进行实验去追逐其他东西。例如，纽约市尝试对其公共项目之一——“家园”（Homebase）进行实验测试，这一项目旨在防止人们流离失所（Buckley, 2010）。符合项目（包括工作培训、咨询和其他援助）条件的人（必须是拖欠房租而且有被赶出去的风险）远多于项目服务所能覆盖的人。因此，纽约市做了合乎逻辑的事，就是去测试项目的功效：他们随即分配了一些人加入这个项目（直到2300万美金花完），另一些相同数目的人则不进入这一项目。这个设计让纽约市弄清了2300万美金花出去之后有多少人从流离失所中被拯救了回来。无论结果是什么，答案都会让纽约市更好地分配资金，如果在这个水平的支出下被拯救的人太少，那么或许资金应该被用于其他方面。相反地，如果大量的人免于流离失所，考虑到无家可归者的社会和经济成本，这个项目应当增加和扩大。无论结果如何，纽约市的民众都得到了更好的服务。

不幸的是，许多纽约的民众和组织并不这样看。他们对“实验”这个鲜活的词语作出了情绪化的反应，并且反对这项旨在让城市更合理地运用资金的对照研究。他们认为这些无家可归者被当作了豚鼠或小白鼠。这些批评者忽略的是，没有人因为实验而得不到服务。无论人群是否被随机分配，都会有相同数量的人接受这个项目的帮助。唯一的不同是通过从控制组收集信息，而不是简单地忽略掉不在项目中的人，纽约市将可以判定这个项目是否起作用！

在这个例子中，对现场实验的误解相当常见。人们似乎不理解在进行现场试验对真实场景中社会援助的作用，我们可以通过什么方法最有效地来使民众得到最大化的帮助。就像国际援助专家艾舍尔·杜夫罗（Esther Duflo）说的：“它看起来并不像是世界观的巨大革新，但是大多数不是经济学家的人无法理解。他们不能理解预算限制。”（Parker, 2010, p.87）我们当我们读到这里的时候，很容易察

觉到杜夫罗话语中包含了些许的沮丧。杜夫罗正在应对我们已经在这本书上论述很多次的事情——对科学家来说显而易见的事情却被外行人完全误解。对杜夫罗来说，在额定的援助预算下，从给定项目中得到服务的人数是一个特定的值。在相同的预算下，另一个更有效的项目能够帮助更多的人。唯一能够断定一个项目是否更有效的方法是进行真正的实验。

或许重构能够对人们有所帮助。杜夫罗的一位对贫困国家进行援助实验的同事说，经常有人对她说："你不应该拿人做实验。"她回复道："好吧，那你就别想知道项目是否有效——那不是实验吗？"（p.87）她在这个问题上回答得相当正确。现行状态——检验它的功效的项目本身也可以被叫作实验，只不过设计得很糟糕！不进行真正实验而运行项目也是一种实验，只不过是没有适当控制的实验！也就是说，这是一种没有控制组的情况！它也是"拿人做实验"！这种重构可能帮助人们消解对寻找什么能够最大限度地帮助人们的客观方法的愚蠢的抵制。

聪明汉斯——神马的故事

用实验控制来排除某种现象的各种其他解释是很有必要的。这种必要性可以通过分析行为科学中一个非常著名的故事来说明。故事的主人公叫聪明汉斯（Clever Hans）——一匹会算术的马。100多年前，一名德国教师向大家展示了一匹马，它的名字叫聪明汉斯，它好像知道如何算术。训练员无论给汉斯出加法、减法还是乘法题，汉斯都能用它的蹄子敲出答案，并且它的回答完全正确。

许多人对于聪明汉斯的表现都感到惊讶和迷惑。难道这匹马真的证明人们低估了这个物种的实际能力吗？人们无疑会有这样的疑问。对汉斯特殊能力的有力见证被德国媒体广泛报道。柏林的一家报社记者写道："这匹会思考的马将会使科学家对许多问题作出长时间的思考。"（Fernald, 1984, p. 30）

这个预言后来被证明是正确的——尽管与记者所期望的有所不同。一组"专家"对汉斯进行了观察，并且证明了它的能力。因此，每个人对此都感到很困惑。这个困惑一直困扰着人们，因为这个现象总是被孤立地观察到，也没有进行任何的控制。但这个谜团很快被一位叫奥斯卡·芬斯特（Oskar Pfungst）的心理学家解开了，他对汉斯的能力进行了系统的研究（Spitz, 1997）。

芬斯特继承了实验设计的优良传统，系统地对动物表演的环境进行操纵，创设了一种“人为”情境（见第7章），这种情境可以用来检验关于马的表现的各种不同说法。在一系列小心谨慎的测试之后，芬斯特发现，这匹马的确具有一种特殊能力，但不是计算能力。事实上，这匹马更像是一位行为科学家，而不是数学家。你看，汉斯是一个非常细心的人类行为的观察者，当它正在敲出答案的时候，它会观察训练员或者出题者的头部。当汉斯接近答案的时候，训练员会下意识地稍微歪一下他的头，然后汉斯就会停下来。芬斯特发现这匹马对视觉线索极其敏感，它能察觉头部的细微动作。于是芬斯特想出了另外一个方法来测试马的能力：就是让不知道答案的提问者向这匹马提问，或者让提问者在马的视线范围以外呈现问题，而在这些情况下，汉斯就失去了它的“数学能力”。

汉斯的例子很好地揭示了仔细区分“对现象的描述”和“对现象的解释”是何等重要。这匹马能够正确敲出训练员呈现给它的数学问题的答案，这是毋庸置疑的，训练员也没有撒谎，而且许多观察者也都证明了这匹马能够做到这一点。问题出现在下一步：即推论这匹马能敲出正确答案是因为它具有数学能力。推断马具有数学能力只是这一现象的一种“假设的解释”。从“马能敲出正确答案”就得出“马具有数学能力”的结论是不符合逻辑的。别忘了，马具有数学能力只是针对马的表现的诸多解释中的一种，而这种解释是可以通过实证方法来检验的。当放在这样一种实验情境下，这个解释就被证伪了。

在芬斯特涉足此事之前，那些见过这匹马的专家们都犯了一个根本性的错误：他们没有想到，对于马的表现还可能存在其他的解释。这些专家认为，只要证明训练员没有撒谎，并且这匹马真的能敲出正确答案，就能够推论出这匹马具有数学能力。然而，芬斯特想得更科学一些，他意识到这只不过是众多可能性中的一种，有必要设立控制条件来区分这些可能性。于是芬斯特设计了一个情境，让训练员站在隔板的后面把问题呈现给这匹马，通过这种方式，芬斯特就可以对两种可能性进行区分：是这匹马真的具有数学能力，还是它能对视觉线索作出反应？如果这匹马真的具有数学能力，让训练员站在隔板后面就不会对马的表现产生任何影响。而如果这匹马是对视觉线索作出反应，那么就会影响马的表现。当后者出现的时候，芬斯特就能够排除“这匹马具有数学能力”这种错误的解释（Splitz, 1997）。

这里可以同第3章中讨论过的节省原则联系起来，所谓的节省原则就是说，当两种理论拥有同样的解释效力时，我们倾向于选择那个比较简单的理论（涉及较少的概念和概念之间的关系）。此处有两种理论，一种认为这匹马具有数学能力，另一种则认为这匹马是在辨别

行为线索，这两种理论在节省原则上的差异是很大的。后者不需要对先前任何心理学和大脑方面的理论作出大幅度调整，它只需要我们将“马对行为线索具有敏感性”的看法稍加调整即可（现在已经广为人知）。而前一种认为马真的能学习算术的理论，则需要我们修改进化论、认知科学、比较心理学和脑科学中的很多概念。这可是相当麻烦的，因为它与其他这些科学缺乏一致性，因此如果它是真的，就需要我们更改这些科学中的很多概念才行（我们将会在第8章讨论所谓的关联原则）。

20世纪90年代的聪明汉斯

聪明汉斯的故事只是一个历史案例，很多年来，在研究方法课上，这个例子都被用来说明实验控制的必要性。没有人认为聪明汉斯的例子会再次出现，但却真的出现了。在20世纪90年代初，全世界的研究者们都在惊恐中观望，就像用慢镜头的方式观察一场车祸一样，眼看着现代版的聪明汉斯的悲剧又一次展现在他们眼前。

自闭症是一种严重的发展性障碍，其表现是社交缺陷、语言发展的滞后及异常，以及活动和兴趣范围狭窄等（Baron-Cohen, 2005）。许多自闭症患儿从外表看起来都很正常，只是极度缺乏与人的交流，这让家长们很难接受。因此，20世纪80年代末期和90年代初期，在澳大利亚有人发明了一种技术，能让自闭的孩子从不说话到自由交流，很难想象这些自闭症患儿的家长们听到这个消息时该是多么激动。这种能让自闭症患者与人交流的技术被称为“辅助沟通疗法”，被一些很有知名度的媒体，如《60分钟》（60 Minutes）、《大观》（Parade）杂志和《华盛顿邮报》（Washington Post）等拿来大肆宣扬（见Lilienfeld et al., 2010；Offit, 2008; Twachtman-Cullen, 1997），据此技术的发明者称，自闭症患者以及其他因发展不良导致言语缺失的儿童，只要把手和胳膊放在这台善解人意的“辅助器”上，就可以在其辅助下，在键盘上敲出相当有文采的句子来。自闭的孩子从之前有限的语言行为到能够交流表达，这种惊人的表现无疑给沮丧的家长们带来了无限希望。这个发明者还宣称，这种技术对于那些有严重智力障碍的失语儿童也同样有效。

尽管家长们的激动心情是可以理解的，但专业人员的轻信盲从就让人不能原谅了。更为糟糕的是，在没有进行控制实验的研究之前，这些媒体节目就开始向抱有无限期望的家长们大肆宣扬这种辅助沟通疗法多么有效。要是这些专业人员在实验控制原则方面受过哪怕一丁点儿训练，他们就能立刻看出这不过是“聪明汉斯”事件的翻版。那些

辅助器可以说是一个永远关注孩子成功的、富有同情心的“人”，在辅助过程中有许多机会有意或无意地指导孩子触碰键盘上的按键。另外一项观察发现，孩子们有时即使不看键盘也能打出复杂的信息，这说明辅助器给了孩子某种暗示。甚至连没学过字母的孩子也能用英语创作出优美的散文。例如，据说一个小孩可以敲出“我是一个奴隶还是自由人？我是身陷囹圄还是被看做友好而理性的精灵？”（Offit, 2008, p.7）。

许多有控制的研究报告称，他们通过适当的实验控制检验了这种辅助沟通疗法。每项研究都明确地说明了同样一件事：自闭症患儿的表现依赖于辅助器发出的不易被觉察的提示（Jacobson, Foxx, & Mulick, 2004; Offit, 2008; Spitz, 1997; Wegner, Fuller, & Sparrow, 2003）。在这些研究中使用的控制方法与聪明汉斯的经典案例是相似的。研究人员设置了一种实验情境，给孩子和辅助器各自呈现一个物体的图案，但是他们彼此看不到呈现给对方的图案是什么。当孩子和辅助器看到的是相同图案的时候，孩子能正确地打出图案的名字；但是当孩子和辅助器看到的图案不同时，孩子打出的是辅助器看到的图案的名字，而不是孩子自己看到的那个图案。因此，答案是由辅助器而不是孩子决定的。

实验结论是，辅助沟通疗法只不过是一种“聪明汉斯”现象，绝非治疗方法上的重大突破，也没有给研究人员带来任何欣喜。但悲剧后面紧跟着更大的悲剧。在一些治疗中心，有当事人在接受辅助器帮助的沟通过程中，讲出过去他们曾受到父亲或母亲的性虐待（Offit, 2008）。于是这些孩子们被迫从家里搬出来，直到这场指控被证明是毫无根据之后，孩子们才被接回来。

由于这些研究结果，专家的意见终于穿透媒体的喧闹浮出水面。重要的是，大家越发认识到，这些缺乏实证基础的疗法并非无害（哦，它有作用，那么它要是没有作用呢？），将未经证实的疗法投入使用是要付出代价的。

俄亥俄州立大学儿科及心理学教授詹姆斯·姆里克（见Mulick, Jacobson, & Kobe, 1993）指出了这种教育手段风行一时所付出的代价：

> 如果没有对辅助沟通疗法的大力宣传，我们可能就会把更多的人力和金钱用于发展基于更有实证基础的、更可行的长远策略，来解决困扰儿童的这一问题。辅助沟通疗法的支持者为研究和专业文献所带来的理论上的混乱，对能力缺陷及其成因方面知识的积累造成了极大的损害……将辅助沟通疗法与其他成功治愈

> 残疾人的非语言交流系统混为一谈，会使真正有效的方法也失去公众的支持……根据我们的经验，残疾人能够成为他们家庭和社区里有价值的成员，他们无需求助于神奇的治疗方法。他们可以寻求现有的有效帮助，这种帮助是有科学意义的。受过科学训练且富有同情心的专业人员的努力胜过所有流行的治疗方法，而且始终如此。治疗的进步和对于治疗的理解是建立在严格的训练、精确的科学标准以及对各种治疗理论的客观证明之上的。（pp. 278-279）

上述这个例子再次证明，仅仅相信见证叙述或者认为流行的治疗方法和伪科学无害，最终会带来危害（见第4章）。由此，我们还能发现，当我们想要正确解释某种行为的时候，实验控制和操纵是不可替代的。

这里需要再次强调一下节省原则。自闭症儿童严重的语言障碍居然能够通过一种“神奇子弹”式（见第9章）的干预方法得到治愈，而这种干预方法推翻了几十年来关于自闭症儿童的认知、神经心理和脑特征的研究成果（Baron-Cohen, 2005; Oberman & Ramachandran, 2007; Rajendran & Mitchell, 2007; Tager-Flusberg, 2007; Wellman, Fang, & Peterson, 2001）。这需要我们修改很多关于认知和神经科学方面已取得的知识。辅助沟通疗法的现状表明，它与其他科学研究成果没有关联性和一致性（见第8章）。

最后，辅助沟通疗法说明了早先聪明的汉斯案例中论述的事情：谨慎地区分描述现象和解释现象的重要性。“辅助沟通”这一术语不是对辅助器和孩子之间所发生的一切的中性描述。相反，它假定了一个理论结果——沟通实际上已经发生了并且是在辅助器的帮助下提升了。但这就是需要被要证明的事情。我们所知的是孩子在敲键。或许如果最初被说成是“意想不到的敲击”，那么事情就会被更理性地处理。需要判断的是“意想不到的敲击”是不是真正的沟通。如果草率地用理论（这就代表了沟通）来给一个现象（按键敲击）加上标签，对实际操作者来说，意识到需要更进一步的调查研究去判定理论能否被证实可能就变得更加困难。

不仅仅是心理学，其他领域也与草率地用理论来对现象进行标签化的问题作斗争。法律系统还在用“婴儿颤栗综合征”这一术语，实际上美国儿科协会已经建议舍弃此用语。这一问题酷似我们论述过的聪明汉斯和辅助沟通的例子。“婴儿颤栗综合征”这一术语是一个理论，有头部创伤的孩子为什么会有这种外显症状。而这一现象是头部创伤自身的特征。创伤的精确描述是通过我们拥有的任何理论来解释创伤是怎样发生的。但今天我们已经知道曾经标准化的术语是误导，法律

系统仍在慢慢经历这种术语变革（Tuerkheimer, 2010）。

交通安全工程师也感觉交通“事故”（accident）这个词带有太多的理论了。事故这个词意味着随机性、不可预测性和运气——纯粹的偶然事件。安全工程师非常清楚汽车交通事故的风险和许多非随机性和可预测的行为之间存在很强的统计学关系。工程师想到像圣路易红雀队投手乔希·汉考克（Josh Hancock）的例子，他租用的越野车与一个停在高速公路上的打着双闪的卡车相撞（Vanderbilt, 2008）。但我们想到汉考克超速（一个巨大的风险因素）、酒精摄入量超过法律限制的两倍（一个巨大的风险因素）、在撞车时正在打电话（一个巨大的风险因素），那么我们把这次撞车说成随机的和不可预测的就大错特错了。他刚刚两天前还和一辆越野车相撞（Vanderbilt, 2008）。把它称为“事故”传递了一种随机性和不可预测性的理论，但是当肆意不计后果的特定行为在这个案例中出现时，随机性和不可预测性似乎并不正确。对事件的描述应当是——汽车相撞。作为一种理论，事故看起来并不正确。

对变量分开考察：特殊条件

戈德伯格与糙皮病的例子给我们上了重要的一课，对于我们澄清有关科学进步的一些错误概念有很大的帮助，尤其是当其运用到心理学中的时候。世界上发生的任何事情通常都与其他许多因素有关联。为了对许多同时发生的事件所造成的因果影响分别进行考察，我们必须创设一些通常情况下不会出现的条件。科学实验将世界上原有的相关分割开来，以此来使单一变量的影响显现出来。

心理学家采取的也是同样的方法：通过操纵和控制来分离变量。例如，认知心理学家们对阅读的过程很感兴趣，他们对促进或阻碍文字识别的因素进行了研究。毫无疑问，他们发现较长的单词比较短的单词更难识别。乍一看，我们会认为单词长度的影响是很容易测量的：简单地设置两组单词，一组长的，一组短的，然后测量两组读者识别速度的差异。不幸的是，事情远没有那么简单。长度较长的词，其使用频率可能也较低，而使用频率本身也会影响识别。因此，长词与短词之间的任何差别都可能是由于长度、使用频率或两个因素共同作用而造成的。为了明确到底词的长度能否独立地对词的识别造成影响，研究者必须创造一些特殊的词，它们的长度与使用频率不是同时变化的。

与之类似，戈德伯格之所以能够作出强有力的原因推断，是由于

他设置了一组非自然发生的特殊条件（想一下他的一个实验操纵是要被试吃下人体的排泄物，这是何等地不自然”啊！）。回想一下奥斯卡·芬斯特设置的一些测试“聪明汉斯”的实验条件，其中包括一些提问者也不知道答案。那些仅仅观察马在自然条件下（提问者知道答案）回答问题的人，非但永远不可能发现那匹马是如何做到这一切的，反而会得出错误的结论，认为那匹马真的具有数学知识。

同样，在检验“辅助沟通疗法”的疗效时，研究者也必须设计一些特殊的条件。呈现给辅助器和儿童的刺激必须分离，这样任何一方都不知道呈现给对方的刺激是什么。为了测试某种现象的不同假设，这类不同寻常的条件是很必要的。

心理学上的很多经典实验都需要将现实世界的自然关系分开考察，通过这样一种逻辑，就能看出哪个变量是决定因素。心理学家哈里·哈洛（Harry Harlow）的著名实验（Harlow, 1958; Harlow & Suomi, 1970）就是个很好的例子。哈洛想要测试一种关于亲子依恋的假设：依恋的产生是由于母亲为婴儿提供食物。然而，问题是母亲提供的不仅仅是食物（还有舒适、温暖、爱抚以及刺激等）。哈洛创设了一种条件，在这种条件下只有一个变量与依恋有关——他让刚出生的短尾猴只能在“人造的”母亲之间选择，并测查了小猴子在这种条件下的行为。例如，他发现，小猴子喜欢厚绒布做成的“母亲”所提供的接触舒适感，甚于喜欢铁丝网做成的“母亲”。出生两周之后，小猴子更喜欢冰冷的厚绒布“母亲”，而不是温暖的铁丝“母亲”，这说明接触上的舒适感比温暖更吸引小猴子。最后，哈洛还发现，即使当食物仅来自于铁丝“母亲”的时候，小猴子仍然更喜欢厚绒布母亲。因此，“依恋仅是由于母亲提供食物”的这种假设是错误的。正是因为哈洛能够对现实世界里同时发生的变量分开进行考察，才会有这样的发现。

创设特殊条件来验证是否存在真正的因果关系，这种方法可以防止错误观念像病毒一样侵袭我们（Stanovich, 2004, 2009, 2011）。让我们看一下关于治疗性触摸的案例，治疗性触摸是在20世纪90年代北美地区十分流行的一种护理方式。使用治疗性触摸法的医生按摩的不是病人的身体，而是病人身上所谓的“能量区”。也就是说，医生的手在病人身体上方游移，但不做真正的按摩。

医生说这是在“感觉”病人的能量区。你会发现，这种感应能量区的能力可以通过创设类似于“聪明汉斯”和“辅助沟通疗法”中的特殊条件来进行验证。也就是说，测试这些医生在看不见的情况下，是否还能感觉出他们的手正接近人的身体。研究结果与聪明汉斯和辅助沟通疗法的案例一样，当视线被挡住之后，这种对距离的感觉能力和随机

水平差不多（Hines, 2003; Shermer, 2005）。这个例子解释了这一章所提到的一点——真实验的逻辑十分直观，小孩都能明白。这是因为，已经发表了的一个实验证明治疗性触摸是无效的，这一实验已经作为学校科研项目完成了（Dacey, 2008）。

简而言之，科学家们用创设特殊条件的方法来验证某种现象的假设是十分必要的。仅观察自然情境还远远不够，人们对下落的和移动的物体观察了几个世纪，却没有人得出关于运动和重力的正确原理和规律。直到伽利略和其他科学家们通过创设人工的条件来观察物体的运动之后，才得到了正确的运动规律。在伽利略的时代，几乎没有人看到过光滑的铜球从光滑的斜面上滚下来。世界上有很多运动发生，但这种运动却非常罕见。这是一种非常规的情境，和其他类似情境一样，使我们第一次得出运动和重力的定律。说到运动定律，在本章最开始的时候，你不是做过一个小测验吗？

直觉物理学

本章开头出现的三个问题实际上是引自约翰·霍普金斯大学的心理学家迈克尔·麦科劳斯基（Michael McCloskey）的一本书。麦克科劳斯基研究的主题被他自己称之为“直觉物理学”。所谓直觉物理学，就是普通人对物体运动的观念。有趣的是，这些观念通常与物体运动的实际情况恰恰相反（Bloom & Weisberg, 2007; Riener, Proffitt, & Salthouse, 2005）。例如，在第一个问题里，当细绳被剪断后，小球会向与细绳垂直的方向直着飞出去（即圆的切线）。麦克科劳斯基发现三分之一的大学生都回答错了，他们认为小球会沿抛物线飞出去。当麦克科劳斯基的被试被问到类似于轰炸机飞行员的那个问题时，有大约一半的人认为应在目标的正上方投掷炸弹，这就表现出他们不理解物体的初始运动决定其后来的运动轨迹，实际上应该在飞机到达目标之前五英里的地方投弹。被试的错误不是因为问题的抽象性质所导致的。当要求被试从房间的一头走到另一头，在走的时候把一个高尔夫球丢在地板上的一个位置时，超过一半人的表现说明，他们不知道高尔夫球下落的时候还会继续向前运动。最后一道题，许多人不知道从步枪射出的子弹落地的时间与子弹垂直落到地面的时间是相同的。

你可以算一下自己在这个小测验中的成绩如何。如果最近你没有上过物理课的话，那么你很有可能至少会错一道题。“物理课！”你可能会提出抗议，“我最近当然没上过物理课，这个测验不公平！”但是请等一下，你为什么需要上物理课才知道这些题目的答案呢？从小到大，你肯定无数次地见过下落的物体。你看到过它们在自然情境中下

落的过程。每天你都能看见运动的物体，你看到的是它们“自然发生”的状态。你当然不能说你对于物体运动毫无经验。

当然，你没见过类似子弹的这种运动。但是我们中的大多数人都见过孩子放开旋转的物体，并且多数人也都见过物体从飞机上落下来。此外，很难说你没见过这些真实的情境。既然你有这么多年关于物体运动和下落的经验，当和真实情境略有不同的时候，为什么你不能准确地预测会发生什么呢？

麦克科劳斯基的研究很好地说明，理解科学家这一做法有多么重要。尽管人们有大量关于物体运动和下落的经验，但对于运动的直觉理论都是相当不靠谱的。我们需要明白的是，外行人观念的不准确是因为他的观察是“自然的”，而不是像科学家那样进行实验控制。因此，如果你在本章开头的测验中错了一道题，不要觉得是自己无知或知识匮乏。要知道几个世纪以前，这个世界上一些伟大人物观察下落的物体后得出的有关运动的物理知识不比现代高中二年级的学生准确到哪去。在《科学美国人》（Scientific American）杂志上的一篇文章中，麦克科劳斯基指出，他观察过的被试中，有很多人都对物体运动持有一种错误的观念，并且这些错误的观念与在牛顿之前三个世纪的理念不谋而合。麦克科劳斯基的当代被试和中世纪哲学家有共通之处：两组人在现实世界里都有很多有关物体运动的经验，但是没有人特意创设一种条件，进行科学的操纵、控制和比较。

直觉心理学

哲学家保罗·丘奇兰德（Churchland, 1988）曾指出，如果我们关于物体运动的直觉（或世俗）理论都是不准确的，那么，也很难相信我们在人类行为这类更为复杂领域中的世俗理论会是正确的：

> 我们最初关于运动的世俗理论是相当混乱的，而且最终将会被更成熟的理论完全取代。早期我们关于宇宙结构和活动的世俗理论也十分离谱，它们之所以依然存留下来，只不过是作为一些历史教训，提醒我们自己可以荒谬到什么程度。我们关于火的本质、生命本质的世俗理论也都是十分荒唐的。由于我们大部分的世俗理论都被推翻了，所以你可以一直列举下去……但是与刚才列出的内容相比，人类的心智活动是一种更复杂和难以理解的现象。目前为止才算有了一些准确的认识，而当我们在其他方面都犯了错误的时候，想要在一开始就能正确地认识心理学知识，简直就是天方夜谭（p. 46）。

当我们审视有关人类行为理论的文献时，会发现丘奇兰德的思考是对的。研究文献警示我们，个人经验并不能为抵御有关心理学的错误信念提供保证。行为经济学家丹·艾瑞里（Dan Ariely, 2008）给我们讲述了他的故事——在他18岁时发生的一场事故导致他全身70%的面积烧伤。他描述了几个月的后续处理，处理时，绷带的快速移除给他带来了极大的痛苦。被护士所信奉的理论是快速除去绷带（导致锐痛）比缓慢的去除绷带（持久但不那么强烈地疼痛）更好。在离开医院并且开始他心理学求学之路时，艾瑞里做实验去测试护士的理念。令他吃惊的是，艾瑞里发现，更慢的过程——更低的疼痛感、更长的持续时间——会降低疼痛的感觉。他说："当我结束实验的时候，我意识到烧伤科的护士都很友善、大方，在去除绷带方面经验丰富，但是她们还是缺乏帮助病人将疼痛最小化的正确理念。我在想她们的经验是如此之丰富，怎么又能错得如此离谱呢？"（p.16）更多的研究表明，即使是临床经验丰富的医师，对其他人身上关于疼痛强度的直觉判断是不正确的（Tait, 2009, & Kalauokalani, 2009）。

正如第4章中所讨论的，依赖见证、个案和"常识"常常会使我们忽略控制组对于检验日常观察得出的结论是否准确的必要性。例如，丁费尔德（Dingfelder, 2006）说过，很多医学专家都认为他们不应该建议抽动性秽语症的患者（见第2章）抑制他们的抽搐（非自主性的语言表达）。医生们相信这会引起一种所谓的回弹效应——抑制后出现更高频率的抽搐。这一观念是基于日常观察而非控制实验。当进行了正确的实验（对比一段时间压抑和没有压抑后的抽搐次数），结果发现所有的抽搐抑制都没有表现出"回弹"效应。

在第1章，我们证明了有关人类行为的许多常识是错误的，这不过是个小的例证而已。例如，没有证据显示有宗教信仰的人比没有宗教信仰的人更无私（Paloutzian & Park, 2005）。研究显示，笃信宗教的程度与参加慈善活动、帮助贫困的人或是不欺骗其他人这些行为之间没有直接关系。

错误的直觉理论不仅限于心理学，它们还盛行于体育界和健身界。例如，定量分析显示，在足球运动中（从高中生到专业球员所有级别），当其他的队伍在中场时，大多数教练相信通过努力争取第四次进攻机会能够增加赢球的可能性（Moskowitz & Wertheim, 2011）。类似的分析显示，总的来说，教练应该更少地踢进攻球而更多使用弃球战术。统计数据证明，如果教练在这些方面重新调整他们的策略，他们能够赢得更多的比赛（Moskowitz & Wertheim, 2011）。现在，教练有一堆理由忽略统计上的建议（例如害怕事后被批评），但是这些理由对球迷不适用。虽然如此，球迷有不正确的直

觉理论：教练是对的。

同理，在健身领域也有很多未经证实的世俗信念。有很多运动员和健身爱好者认为在运动开始前进行拉伸能够防止运动中受伤。但是证据显示的并不是那样（Bernstein, 2009）。类似地，大多数的跑步者知道10%法则：如果锻炼时增加跑步距离，保证每周增加的距离不超过10%就能够避免受伤。但问题是“10%法则”未经研究证实（Kolata, 2011）。最后，当棒球运动员在击球区进行击球练习时，许多球员在他们棒球上套一些有重量的圈状物。研究显示这些重物弊大于利——但这无法说服球员不用这一招术（Wolff, 2011）。

有关人类行为的错误观念可以产生非常实际的后果。基思和拜因斯（Keith & Beins, 2008）提到，在学生中，对于手机和开车的典型观点是这样的：“通话不影响我开车”和“打手机可以防止我睡着”。学生们似乎完全不知道在开车时使用手机（即使无需手持）会严重影响专注和注意（Kunar, Carter, Cohen, & Horowitz, 2008; Strayer & Drews, 2007），而这会导致事故和死亡（Conkle & West, 2008; McEvoy, et al., 2005; Parker-Pope, 2009）。

假如列出所有错误的世俗观念，那么这个清单会很长。例如，很多人认为“月亮盈亏会影响人的行为”，事实并非如此（见Foster & Roenneberg, 2008；Lilienfeld et al., 2010）；一些人认为“性格互补的人相互吸引”，他们也错了（Gaunt, 2006; Hitsch, Hortacsu, & Ariely, 2010; Reis, Maniaci, Capraiello, Eastwick, & Finkel, 2011）；有些人相信在做选择题的时候，千万不要更改答案，他们错了（Kruger et al., 2005）；有些人相信做祷告让人更健康，其实并不能（Benson et al., 2006）；有些人相信“亲生厌，熟生蔑”，实际没有这回事（Claypool, Hall, Mackie, & Garcia-Marques, 2008; Zebrowitz, White, & Wieneke, 2008）。这种例子不胜枚举（见Lilienfeld et al., 2010）。

人类关于行为的直觉理论是有缺陷的，这就说明了为什么我们的心理学研究需要实验控制。只有这样，我们才能把关于人类行为的粗浅概念上升为准确的科学概念和体系。

小结

实验方法的核心就是操纵与控制，这就是为什么实验比相关研究能够作出更强的因果推断。在相关研究中，研究者仅仅观察两个变量的自然变动是否显示某种联系，而在真实验中，研究者要对被假设为原因的变量进行操纵，通过实验控制和随机分配来保持其他所有变量不变，然后再来看这个假设变量是否会产生影响。这种方法排除了相关研究中出现的第三变量的问题。第三变量出现的原因是，在自然情境下，很多不同的事物都是相互联系的。实验方法就是用来分开考察这些自然存在的关联。它之所以能实现这一目的，是因为它以操纵一个变量（被假设是原因的变量）的方式分离出该变量，并保持其他所有变量不变。但是，为了区分这些自然的关联，科学家们经常要创设自然世界里不会出现的特殊条件。

Chapter 7
不像是真实生活的心理学实验与“人为性”批评

前两章讲述了实验逻辑的原则，现在我们可以思考一下心理学经常面对的一些批评。比如，很多人认为科学实验没有价值，因为它是人为发生的，和“真实的生活”不一样。我们将对这一观点进行详细探讨。由于心理学实验常常遭到类似的批评，因此理解这种批评的不合理之处，将有助于我们更好地了解心理学。

为什么自然性并非总是必要的

从第6章的内容中，我们已经可以清楚地看到为什么这种批评是不合理的。正如第6章所述，科学实验的人为性并不是一种缺点，事实上，正是它使得科学方法具备了一种奇特的力量，可以让我们对世界进行解释。与人们通常所相信的观点不同，科学实验的人为性并不是偶然的疏忽，而是科学家故意为之。科学家之所以专门设置一些非自然发生的条件，是因为只有这样才可以将决定事件发生的许多相关变量区分开来。用第6章的话来说，科学家设定特殊的条件是为了分离变量。

有时候，必要条件已经在自然状态中存在，比如斯诺和霍乱病的例子。但这种情况并不经常出现。科学家必须用新异的、甚至有时比较奇怪的方法操控事件，比如戈德伯格和糙皮病的例子。很多时候，这些操作无法在自然环境中完成，于是科学家必须把所要研究的现象转移到实验室中，以便实施更精确的控制。例如，在有关“重力和运动”的早期研究中，使用了一些特制的物体，其目的就是为了创造一些特殊条件，以便观察物体运动。因此，为了分析一种现象，经常需要创设非自然的极端条件。

事实上，如果科学家完全禁锢在“自然”条件下观察，那么一些现象就不可能被发现。探索物质本质特征的物理学家们建造巨大的加速器来诱发基本粒子之间的碰撞。碰撞中产生的一些副产物是存在时间不到十亿分之一秒的新粒子。然而，这些新粒子的属性却有助于解释原子结构理论。许多新粒子在世界上一般是不存在的，即使存在，我们在自然状况下也没有机会观察到它们。因此，几乎没有人质疑物理学家们的研究方式。为了对宇宙有更深刻的理解，即使采用一些不常见的、甚至是怪异的方法，也是合情合理的。但不知为什么，物理学家用起来合理的方法，心理学家使用起来就常被认为是不合理的。

许多心理学家在向外行人展示关于某一行为的实验证据之后，都听到过这样的叹息：“可惜这不是现实的生活！”这种评论反映了人们的一种观念：在实验室研究人类心理是件奇怪的事。这种拒斥还包含了一种假设：知识的获得只能通过自然条件下的研究。

心理学家使用的许多技术在公众看来是怪异的，很多人都不知道这些技术并非心理学领域所独有，只不过心理学家把这些科学方法应用到人类行为的研究上而已。禁锢于真实生活条件会妨碍我们发现许多新事物。例如，生物反馈技术现在被广泛应用于各种领域，比如用于控制周期性偏头痛和紧张性头痛、治疗高血压，以及放松训练（de Charms et al., 2005; Maizels, 2005）。研究表明，如果通过视觉或听觉的反馈能够监测到体内正在进行的生理过程，那么人类就能学会在一定程度上控制这些过程。这项研究促进了上述生物反馈技术的发展。当然，因为人类本身并不具备通过外部反馈来监测自身生理功能的能力，所以，如果不是在特殊的实验室条件下，人们将很难发现人类有能力控制自己的生理过程。自然条件下的观察是永远无法发现这一点的。

对“随机取样”的误解

然而有时候，类似“这不是真实的生活”的抱怨源于对心理学实验研究目的的另一种误解，产生这种误解的原因是非常容易理解的。媒体的宣传使许多人对调查研究开始熟悉起来，特别是选举中的民意调查。现在人们对选举投票的一些重要特征越来越了解。具体而言，为了保证民意测验的准确性，媒体对随机取样、样本代表性等概念更加关注。这种关注导致许多人错误地认为，随机取样和代表性是所有心理学调查研究的必要条件。因为心理学研究很少使用随机的被试样本，如果根据外行人所相信的随机取样标准，那么许多心理学的研究成果都会遭到诋毁，那些批评心理学研究无法反映真实生活因而是无效的论点也会受到强化。

但只要想一下其他科学的情况，就很容易理解这种想法的荒谬。化学家从没尝试过抽取化合物的随机样本，生物学家也不曾用细胞或组织的随机样本进行实验。用于医学研究的老鼠和猴子也不能完全代表其物种。而这些研究都是在与这些动物生活的自然环境完全不同的实验室中进行的。事实上，这些条件通常很独特。然而，所有这些研究得到的结果都可以帮助我们理解人体生物学。大部分心理学研究也是同样的道理。并非每一个心理学调查研究都需要使用随机样本。因此，我们在此需要强调的重点是：随机取样和随机分配（见第6章）不是一回事情。

随机分配和随机样本的区别

随机分配和随机取样两个词里都包含“随机”，因此许多人以为它们所指的是一回事。事实上，它们是两个非常不同的概念，唯一相似之处在于它们都采用了随机生成数字这一点，然而其目的却大相径庭。

随机取样涉及的是如何选择被试进行研究。如前所述，并不是所有的研究都要求随机取样，但当它成为必要条件时（例如在调查研究、市场调查或是选举时的民意调查中），我们则需要用一种方法从总体中抽取一个样本，这种方法要确保总体中的每一个成员都有同等机会被选为样本，被抽中的样本就成为随后调查研究中的被试。有一点非常重要，这种随机抽样的调查研究既可能是相关研究，也可能是一个真实验。只有使用了随机分配的方式，才有可能成为一个真实验。

随机分配是真实验所必需的条件。实验人员将被试分为实验组和控制组，当每一名被试被分到实验组的机会和被分到控制组的机会相等时，就实现了随机分配。为了达到这一点，常会用到像掷硬币这样的随机化手段（更常用的是一种特殊的随机化数字表格）——因为它在给被试分组时没有任何偏向。

随机分配和随机取样不是一回事，牢记这一点的最好方法是弄清楚四种组合：非随机分配的非随机样本，随机分配的非随机样本，非随机分配的随机样本，以及随机分配的随机样本。大部分心理学实验没有使用随机样本，因为没有这个必要。正如第8章将讲到的，研究可以检验理论，我们所需要的只是一个方便取得的样本。如果一个研究中使用了随机分配的方法，那么它就是一项真实验，如果没有使用，那么它是一项相关调查。许多使用随机取样的研究没有使用随机分配，那是因为它们只是调查性研究，旨在寻找关联——也就是说，这些研究属于相关调查研究。

理论研究和应用研究的异同

弗吉尼亚大学心理学家道格拉斯·穆克（Douglas Mook）阐述了不同类型的研究要求的不同类型的预测。许多应用研究的目的是把研究结果直接与生活中的特殊情境联系起来。选举投票中的民意测验就是应用研究的一个例子。研究目的是预测一个特定情境下的特定行为，在这个例子中，就是选举日的投票结果。由于研究结果是要直接应用于现实的，因此样本的随机性和情境的代表性问题就显得尤为重要。

然而，把应用型心理学研究看做典型的心理学研究是错误的。心理学（或其他学科，就这一点来说也是如此）的大部分研究都有着不同于应用的目的。它们的目的都是为了发展理论。大多数研究的结果只能间接通过理论修改而被应用，这些理论与其他科学规律共同应用于一些实践性问题（Nickerson, 1999）。简而言之，大部分理论研究追求的是对心理过程的理论验证，而不是把研究结果推广到现实中的某一特殊情境中去。

主要目的为理论验证的研究通常被称为“基础研究”。应用研究的目的是把数据直接应用于现实生活，但是基础研究则专注于理论验证。然而，仅仅根据某项研究是否有实践性应用来区分基础研究和应用研究，很可能会产生错误，因为这一差别常常会随着时间的增长而逐渐消失。应用研究的结果会很快得到应用。但是没有什么能比普遍的、准确的理论更具有实用性了。尽管很多科学家进行理论或实证研究的初衷并非解决具体的实践性问题，但他们发展出的科学理论或研究结果最终都解决了现实世界的许多问题。这样的例子在科学史上不胜枚举。

历史一再证明，（通过让科学家解决特殊的实践性问题而）试图控制科学发展方向只能阻碍发展进程而非促进。具有讽刺意味的是，急于让科学家们解决实际问题，而不让其考虑“其他事情”（基础研究）的做法被证明是最不切实际和目光短浅的，因为通向实际应用的道路充满着不可预知性。为了研究关节炎，得克萨斯西南大学医药研究中心的一组研究人员试图通过遗传的方式培养一批患有关节炎的老鼠。出乎意料的是，这些老鼠同时也出现了类似溃疡性肠炎的肠感染（Fackelman, 1996）。科学家们从此拥有了研究人类疾病的动物模型”（Fackelman, 1996, p. 302）。无论这些科学家是否在关节炎（原本想研究的问题）上取得了进展，现在看来他们似乎在溃疡性肠炎和克罗恩病（即节段型肠炎）的治疗上作出了巨大的贡献。这种间接性关联的取得在科学中比比皆是。辉瑞制药（Pfizer）发明伟哥的时候其实是在寻找一种治疗心脏病的药物（Gladwell, 2010）。

基础研究和应用之间的这种间接联系让人难以理解。基础研究似乎与现实问题相距甚远，基础研究也因此容易受到讥讽。20世纪70到80年代，美国参议员威廉·普罗克斯迈尔（William Proxmire）选出了一些题目听起来比较奇怪的基础研究，并认为这是政府资金被浪费的证据（Benson, 2006a; Munro, 2010）。但是随着时间的流逝，沦为笑柄的是参议员普罗克斯迈尔，而不是研究者。普罗克斯迈尔参议员因为名字听上去愚蠢而选出的研究（例如“猴子为什么紧咬牙关”），被反复证明引领了理论进展或实践应用。例如，猴子紧咬牙关的研究

使得紧张这个概念操作化。这对政府机构客观评估人们在密闭空间（如外太空或潜水艇）进行作业时的紧张程度有极大帮助（Benson, 2006a）。

被参议员普罗克斯迈尔嘲笑的研究最终证实为真正有用的，这一幕在2008年美国总统竞选时再次上演。候选人约翰·麦凯恩嘲笑蒙大拿州关于熊DNA的研究（Krauss, 2008）。他的竞选搭档莎拉·佩林（Sarah Palin）批评这类研究在法国巴黎进行的“果蝇实验”，认为这无关公众的福祉（Krauss, 2008）。提及这些研究，或许能够成功地迎合人们的观念：科学就是浪费钱，但不得不说他们也太会选了。熊的研究被证明是受联邦濒危物种法案委托，在来自美国地质调查、美国鱼类和野生动植物服务机构、蒙大纳州鱼类和野生动植物及公园服务机构的科学家们的推荐下进行的。所有这些机构都认为，研究者所查明的熊的数量和位置，对于保护濒危动物极为关键。

佩林的选择尤为糟糕，简直具有讽刺性。首先，法国的实验室是受美国农业部资助，因为在加州遭受橄榄果蝇虫害之前几十年，法国曾爆发过同样的虫害（Krauss, 2008）。对美国来说，控制橄榄果蝇虫害就有即时的经济收益。更讽刺的是，佩林的演讲还部分涉及联邦残疾人教育法案，而她自己就有一个存在智力缺陷的孩子。果蝇曾是（并将继续是）基因研究领域一个关键活体组织——该领域与被归入联邦残疾人教育法案保障范围的各类残疾的诊断和治疗直接相关。

自从那次竞选开始，政治家们总是为了表现得与众不同而去炫耀他们对科学运作的误解。俄克拉何马州参议员汤姆·科伯恩（Tom Coburn）抨击了美国国家科学基金的社会科学诸多分支，这一基金为研究对理解经济非常关键的行为经济学工作提供资助，并且把资助给诺贝尔奖得主（Cohen, 2009）。他挑出一些他认为具有“滑稽”标题的科学项目进行批评，包括涉及一些有效的冷冻大鼠精子方法的一个项目。毫无疑问，他通过让选民嘲笑“老鼠精子科学”在俄克拉荷马州捞取了不少政治资本，但是最终沦为笑柄的是他自己。美国国立卫生研究院的主任在听证会上回答了“为什么会有人把钱花在冷冻老鼠精子上”这个问题。弗朗西斯·柯林斯博士（Dr.Francis Collins）解释说：“我们用这些非常有价值的老鼠种系，这些种系代表了人类疾病的特殊模型，比如高血压和心脏病……你只需仅冷冻老鼠的精子，当需要的时候你就能重新创造那只老鼠，这会为你节省一大笔钱。知道如何高效地做这件事是一笔很划算的投资。当然，没人会用心了解做这件事的原因。人们只是觉得这听起来很怪异，感觉像是在浪费钱。”（Boyer, 2010, p.62）科学发现包括把科学的不同领域的发现整合起来，这种联系常常在外行人眼里不那么显而易见。

我们必须意识到，虽然一些研究是为了直接预测某一特殊情境而设计的，但大多数科学研究仍然是用于验证理论的基础研究。怎样把研究结果应用到现实生活中呢？从事应用研究和从事基础研究的研究者们对此有不同的回答。前者会这样回答："直接应用，只要实验情境和将来要应用的情境有相当程度的相似性就可以了。"因此，被试的随机取样和实验情境的代表性都会影响结果的应用。然而，进行理论检验的研究人员会这样认为："研究结果不会直接应用于现实生活，进行理论研究的目的也不是为了将结果用于具体的环境条件。"因此，这类科学家并不关心研究的被试与其他群体有多相似，也不关心实验情境是否反映出某些真实生活的环境。那么，这是否意味着这些研究结果对现实世界没有意义呢？不是的。这些研究结果不直接应用于某一特殊情境，而是应用于理论。这种理论也许在将来的某一天，可以和其他科学规律相结合，共同解决某一特殊问题。

在心理学的一些领域里，这种将理论间接应用于现实生活的研究十分常见。例如，许多年前，手机刚刚面世，许多认知心理学家立即开始担心安全问题——人们边开车边接听手机怎么办？心理学家立即预测手机的使用可能会导致交通事故增多，不仅仅是因为接听电话的时候一只手会离开方向盘，此外，他们还担心接听电话会转移司机的注意力。有一点很重要，我们应该意识到，心理学家提出这些担忧远远早于真正用移动电话来做的实验研究（见Strayer & Drews, 2007; Strayer & Johnston, 2001）。心理学家通过理论预测手机事故问题，而这个例子中的注意力有限加工理论早在几十年前就已经存在了（如，Broadbent, 1958; Kahneman, 1973）。这一信息加工理论是通过大量的实验验证（上百个实验室研究）建立起来的，开车使用手机提供了一个机会，正好可以用这一理论来预测其可能造成的危害。事实也是如此，后来使用移动电话进行研究，结果证实了心理学中注意理论的预测：移动电话的使用确实是引发交通事故的一个原因（Conkle & West, 2008; Insurance Institute for Highway Safety, 2005; Kunar et al., 2008; Levy, Pashler, & Boer, 2006; McEvoy et al., 2005; Redelmeier & Tibshirani 2001; Strayer & Drews, 2007）。

道格拉斯·穆克（Douglas Mook, 1983）就一个例子阐述了心理学中通过实验来验证理论的观点以及间接应用的性质。20世纪30年代，塞里格·海奇特（Selig Hecht）在《普通实验心理学手册》（Handbook of General Experimental Psychology）（Murchison, 1934）里发表了一系列有关视敏度的研究，谈到了暗适应的现象。你可能有过暂时性"失明"的经历，比如当你走进一个漆黑的电影院时。但是，当你在位置上坐了一会儿之后，应该就能注意到椅子、人以及

其他物体慢慢变得可以看见。如果你继续关注这个现象，你会发现视敏度不断升高的这个过程会持续几分钟之久。

这种现象叫做暗适应，它会经历两个阶段：首先是在进入一间漆黑的屋子时，视敏度迅速小幅度地升高，之后缓慢大幅度升高。海奇特把两部分的升高曲线和视网膜上的两种感光细胞联系起来。密集分布在中央窝中心（视网膜的一部分，用于聚光）的视锥细胞，对红光非常敏感。分布在中央窝外围的视杆细胞没有那么密集，而且对红光不是很敏感。海奇特根据这些事实建立了一个理论，即暗适应的最初阶段（视敏度小幅地快速升高）源于视锥细胞的适应，第二阶段（在更长的一段时间内视敏度大幅升高）源于视杆细胞的适应。

穆克（1983）提醒我们考虑一下海奇特的实验环境是完全非自然的。（非随机取样的）被试在暗室里进行反应，根据他们是否察觉到微弱的红色闪光，回答“是，我看得见”或者“不，我看不见”。正常情况下，我们不会在日常生活中对微弱的红光作“是”或“否”的反应。然而由于海奇特并不考虑将自己的研究成果推广到那些在暗室里对红光做“是”或“否”的反应的个体中去，所以现实生活中这种情况是否真的发生过无关紧要。海奇特所关心的是，如何根据实验室中建立的事实来验证相应的理论，从而能解释视觉系统所特有的一些基本过程，如暗适应。他并不关心他的实验情境是否符合现实世界的情况，而是关注是否能完整分离出他想研究的特殊视觉过程。

海奇特的研究发现之所以具有普遍性，并不是因为他的实验情境的性质是人工的或是自然的，而是因为根据这些研究结果可以建立一个有关基本视觉过程的理论，而这个理论可以与许多视觉现象相关联。他的研究揭示了人类视觉系统中各个部分之间的功能关系，而这恰恰是因为他的研究情境经过了人为的精确控制。如果这一理论模型是正确的，那么它应该能广泛地应用于各种情境，可以用来解释许多行为现象，即使所处的情境与发现这一理论的情境完全不同。事实上，海奇特的研究结果通过对理论的影响而产生了间接的应用价值。例如，海奇特的研究结果促进了对夜盲症的治疗，改善了X射线的识别问题（Leibowitz, 1996; Mook, 1982）。更引人注目的是，第二次世界大战期间，英国飞行员在闪电战中等待希特勒轰炸机的夜间袭击时，戴上了红色的飞行眼镜（因为视杆细胞对红光不够敏感，可以保持暗适应；见Mook, 1982）。从在实验室里判断小红点到辨别伦敦上空危险物体的移动，这一鸿沟是由理论跨越的，而不是通过把海奇特的实验室改造成喷气式战斗机得出的。

海奇特的例子表明，心理学发现走向应用已有数十年的历史。在第二次世界大战期间有很多这种应用的例子。例如，在战争开始时，

盟军的海军发现，海军人员识别敌我飞机和舰艇的时间过长（Joyce, 2010）。他们求助于俄亥俄州立大学的实验心理学家塞缪尔·伦肖（Samuel Renshaw），并询问他是否能够找出一种整体识别方法，能够比他们现在的WEFT（机翼、引擎、机身、机尾）识别系统更快。伦肖提出了一个在实验室测试结果较好的整体性识别方法。这一方法在战场上的应用相当成功，伦肖也因为拯救成百上千个生命而被赞颂。

心理学理论的应用

一旦我们明白了大部分研究的目的是发展理论而不是预测具体环境下的事件，以及大部分研究的结果是通过理论间接应用的，而非在具体环境条件下直接应用，那么我们就会顺理成章地发问：究竟心理学中有多少理论可以在现实中得到应用？也就是说，心理学理论的普遍性得到验证了没有？

对于这一点，我们必须承认以往的记录是参差不齐的，但也必须清楚地意识到，这与心理学的多样性息息相关。一些领域中的研究确实在应用方面进展甚微，然而其他一些领域则已经取得了十分瞩目的成绩，通过实验已经推导出了许多具有解释能力以及预测效力的原理。

想想经典条件反射和操作性条件反射原理。这些原理及其详细论述的规律，几乎完全是从非人类被试的实验发展而来的，比如鸽子、老鼠，其实验情境也是高度人为化的实验室环境。然而，这些原理已经成功地用于解决人类各式各样的问题，包括自闭症儿童的治疗、大量事实材料的教学、酗酒和肥胖症的治疗、精神病院的病号管理以及恐惧症的治疗，等等。这仅是一小部分而已。这些应用所依靠的原理之所以能够被准确地提炼出来，是因为在实验室条件下，研究者们能够精确地细化环境刺激和行为之间的关系，而这一点在自然条件下是无法做到的，因为在自然情境下，许多行为之间的关系可能会同时起作用。至于非人类被试的使用，是在许多案例中，从动物的反应得出的理论和规律为我们提供了与人类行为非常相近的数据（Vazire & Gosling, 2003）。人类研究发现，人类的行为规律与从动物行为得出的规律非常相似。当人类疾病治疗方面的每一次医学进步都源于动物研究数据的时候，这些发现也不应该再让我们感到惊奇了。举例来说，动物研究促进了很多领域的发展，包括行为医学、压力缓解、心理治疗、受伤或者残疾人士的康复、研究衰老对记忆力的影响、帮助人们克服神经性肌肉紊乱的方法、了解药物对胎儿发育的影响、交通安全、慢性疼痛的治疗（Gosling, 2001; Kalat, 2007; Michael, 2008; Zimbardo, 2004）。最近，关于狗的研究为了解人类焦虑障碍的基本原理起到了实质性的推动作用（Mineka & Zinbarg, 2006）。

事实上，“这不是真实的生活”的批评被错误地用来诋毁动物研究的成果——这种做法经常是由于政治的缘故。例如，那些为重度污染企业效劳的政客们总是否认致癌因素风险评估报告的有效性，他们的理由是，这些报告是以动物研究为基础的，不能应用到人类风险评估上。然而，一组科学家在1988年进行的一个对23种致癌物质（苯、石棉，等等）的研究中发现，由动物研究计算出来的死亡率与由人类流行病学研究计算的结果非常相近（Finkel, 1996）。

心理学家对知觉过程的研究取得了令人瞩目的进展，从中得出的规律和理论已用于解决各种各样的问题，比如雷达监测系统、街灯照明以及飞机驾驶舱的设计（Durso, Nickerson, Dumais, Lewandowsky, & Perfect, 2007; Swets, Dawes, & Monahan, 2000; Wickens, Lee, Liu, & Gordon-Becker, 2003）。关于衰老对认知的影响，我们已经积累了许多新的认识（Salthouse, 2012），而这些新知识有可能会直接帮助我们设计出让认知丧失者恢复其能力的训练方案（Schaie & Willis,2010）。

判断和决策的心理研究已经应用于医学、教育和经济等领域的决策制定（Adler, 2009;Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz, & Woloshin, 2007; Kahneman, 2011; Stanovich, 2011; Tetlock, 2005; Thaler & Sunstein, 2008; Zweig, 2008）。斯坦利·米尔格拉姆著名的强迫服从实验就被应用于军队的训练。一个激动人心的进展是认知心理学家更多地参与到法律体系中，在这一过程中，取证中的记忆、证据评估和作出判决等方面的问题为检验认知理论的应用性提供了大量机会（Spellman & Busey, 2010; Wargo, 2011; Wells, Memon, & Penrod, 2006）。近几十年来，阅读教学中的理论和实践开始受到认知心理学的影响（Hulme & Snowling, 2011; Presley, 2005; Snowling & Hulme, 2005; Stanovich, 2000）。

简而言之，心理学在很多方面都贴近“真实生活”，只不过很少为大众所知。研究型心理学家已经找到了怎么使人们节省更多的养老金以及如何增加器官捐赠比例（Thaler & Sunstein, 2008）的途径，发现了如何让人们去注射流感疫苗（Price, 2009），创新出了能够减少能源使用的行为模式（Attari, Dekay, Davidson, & Rruine de Bruin, 2010; Todd, 2010），发现了促进屏幕阅读的方法（Chamberlin, 2010），帮助政府增强了人们依法纳税（Hill, 2010），寻找到了节省健康支出的方法（Deangelis, 2010），找到了由来已久的为什么孩子不愿上学的答案（Willingham, 2009），以及提高了投票率（Bryan, Walton, Roger, & Dweck, 2011）。

在有关儿童在法律程序中提供的证词（Bruch & Ceci, 2004）和

受虐儿童所“恢复”的记忆是否准确（Brainerd & Reyna, 2005; McNally & Geraerts, 2009; Moore & Zoellner, 2007）等这些公众争论不休的问题方面，心理学家都提供了重要的科学依据。认知心理学家芭芭拉·特维斯基（Barbara Tversky）研究空间认知，其成果的衍生已经用于设计计算机地图路线发生器、编写DIY组装家具的说明书（Benson, 2006b）。认知和实验心理学博士朱迪·瑟依（Judi See）致力于将心理学应用于军事问题，她以她的职业生涯向我们展示了心理学的应用潜力（See, 2006）。在其多样化而又吸引人的职业生涯里，她评估了全球鹰无人机的监视质量；评估了防毒面具所加载的镜片；帮助B-2飞行员安排睡觉和清醒的时间以抵御任务中的疲劳；评估了在伊拉克使用的手持翻译设备，并且将信号检测论用于爆炸装置的解除（See, 2006）。

美国心理协会（APS）有一个网站，你可以在此网站上看到更多有关心理学知识的实践性应用。这个网站叫做“我们只是人类而已”（We’re Only Human），由沃瑞·赫伯特撰写。许多心理学研究的应用都会在这里讨论：www.psychologicalscience.org/onlyhuman。《科学美国思想》（Scientific American Mind）这本杂志也会报道许多心理学应用。

“大二学生”问题

许多人质疑心理学研究成果的代表性，他们过于关注研究的被试，而不关心实验设计的细节。我们面临这样一个问题，有时人们把它称作“大二学生问题”，即因为大二学生在大量的心理学研究中充当被试，因此这些研究所得出的结果是否具有可推广性受到了质疑。心理学家之所以关心这一问题，是因为它在某些研究领域中的确是个问题。尽管如此，我们还是要正确地看待它，并且应该知道心理学家对这一批评有几种合理的辩解。以下列出了三点。

1. 这种批评不能说明研究结果无效，只是需要更多的研究来证明理论的可推广性。由于我们先前收集了大二学生的数据，即使从其他人群中获得了相反的数据，从而必须对理论作出相应的调整，也只会使理论更加精确，而不会完全否定它。即使在比较极端的情况下，重复的实验没有得出相同的结果，我们也只能说，建立在大二学生数据基础上的理论不够全面，而不能说该理论一定是错误的。

2. 在心理学众多领域里，大二学生问题不构成一个问题，因为所研究的心理过程是非常基本的过程（例如视觉系统），几乎没有人相

信视觉系统的基本构造跟被试样本的人口分布特征有关。美国蒙大拿州和佛罗里达州（或就此而言，阿根廷）的人们在基本的信息加工操作、大脑的功能性组织和视觉系统的性质上非常相近。除此之外，这些人类的特征和一个人的父母是修鞋匠、裁缝抑或是教授没有多大关系。

3. 许多研究结果得到了重复，这使我们确信这些结果在很大程度上可以推广到不同的地理分布中，并且在较小程度上也能推广到具有不同社会经济因素、家庭变量以及早期教育经历的人群中去。50年前的大学生被试样本恐怕基本来自精英团体，如今却完全不同，现在大学生的家庭背景能够代表各阶层的群体。

然而，否认大二学生问题在心理学研究的某些领域里的确是个问题，亦非明智之举。尽管如此，心理学家正在尽力矫正这个问题。例如，发展心理学家几乎都很关注这个问题。在这一领域中，每年都有成百上千的研究人员将众多用大学生被试得出的理论和发现在其他不同年龄的被试身上重新验证。

用不同年龄组的人做被试并不总是能重复用大学生被试得出的结果。要是那样的话，发展心理学就会变得很无聊了。但是一大堆心理学家都致力于在心理学理论中建立一个年龄因素，以证明这个因素的重要性。这一领域的研究也确保了心理学的宏大理论不是只建立在从大学生那里收集的有限数据基础之上。

许多发展过程的研究都是以北美儿童为被试进行的，为了评估这些研究结果的可推广性，发展心理学家也进行跨文化的研究。许多跨文化比较的例证也表明了跨文化的相似性（例如Demetriou et al., 2005; McBride-Chang & Kail, 2002）。但是，也有不少跨文化研究没有显示出与美国大二学生相似的结果（例如Buchtel & Norenzayan, 2009; Henrich, Heine, & Norenzayan, 2010）。但是当出现这些差异时，这些研究仍然提供了一些重要信息，让人们了解到，这些理论和结果会因文化和背景的不同而不同（Buchtel & Norenzayan, 2009; Henrich et al.，2010; Medin & Atran, 2004）。

正如先前提到的，认知心理学的研究成果通过了重复验证。信息加工的许多基本规律在全世界许多实验室中得到验证。人们可能不太知道，如果密歇根大学的一名心理学家获得一项重要的研究成果，那么类似的实验将很快在斯坦福大学、明尼苏达大学、俄亥俄州立大学、剑桥大学、耶鲁大学、多伦多大学等大学进行。通过这种检验，我们将很快知道这项结果是不是由于密歇根州被试的独特性或特殊的实验环境所造成的。

大二学生问题和关于代表性的批评大部分针对的是社会心理学。社会心理学经常用大学生被试在实验室情境中进行研究，并试图建立真实社会情境中的社会交往、群体行为和信息加工等理论（Myers, 2006）。然而，即使在心理学的这一领域，也有证据表明，实验室得出的成果和理论，实际上确实预测出了不同类型的个体在各种情境下的行为。

例如，几年以前，威斯康辛大学的一名心理学家莱昂纳德·伯克维茨（Leonard Berkowitz）证明了所谓的“武器效应”——如果一件武器出现在手边，会使得某个人更容易作出攻击性反应。这个发现源于实验室，是一个无代表性情境的典型例子。由于这一结果是人为情境的诱导产物，因此常被强烈地批评其具有误导性。但事实是这样的，各种实验条件下得出的结果是一致的，用不同的方法测量攻击性所得的结果是一致的，在欧洲和美国获得的结果是一致的，研究儿童和成人的结果是一致的，在实验室之外的现场研究中，被试不知道自己是在参与实验，得出的结果也一样（Berkowitz & Donnerstein, 1982）。研究人员甚至提取出了武器效应背后的认知机制。在语义记忆中，它是一个自动启动的过程（见Meier, Robinson, & Wilkowski, 2007; Wilkowski & Robinson, 2008）。

认知、社会和临床心理学家也研究了人类的各种决策行为。这个研究领域里大部分初始性研究都是在实验室里完成的，以大学生为被试，而且采用高度人为化的任务。然而，从这些研究中得出的决策行为原则在很多非实验室环境中都得到了重现，包括银行家对股票价格的判断、赌场赌博、精神病医生对病人行为的预测、经济市场预测、军事情报分析、全美橄榄球联赛的下注、工程师对修理时间的估计、房地产经纪人对房价的估计、商务决策以及医生的诊断，这些原则现在也被应用于个人理财咨询的实践领域（Adler, 2009; Ariely, 2008; Hilton, 2003; Kahneman, 2011; Stanovich, 2009; Thaler & Sunstein, 2008; Zweig, 2008）。

伯恩鲍姆（Birnbaum, 1999, 2004）用互联网来解决心理学中的大二学生问题。他在实验室里通过互联网招募了一批参与者，并进行了一系列有关决策问题的实验。实验室中得到的结果全部在互联网样本中得以重现，而后者的取样范围要比前者广泛得多——包含来自44个国家的1224名参与者（也见Jaffe, 2005;Skitka & Sargis, 2006）。高斯林等人（Gosling, Simine, Srivastava, & John, 2004）研究了大量互联网被试样本（361703人），并将之与发表过的510个传统样本的被试比较，发现互联网上的被试在性别、社会经济地位、地区和年龄方面有着更广泛的分布。重要的是，他们发现心理学众多研究领域

的研究结果，例如人格理论，用互联网实验和传统方法的研究所得出的结果非常相似。

并非所有的心理学研究成果都能被重复。相反，结果无法重复的实验经常出现，而它们往往比结果可以重复的实验更具指导意义。但是，在认知心理学中，重复实验的失败几乎很少是由被试的独特性造成的。相反，大部分是源于实验刺激和方法的细微差异。通过仔细分析要产生一个现象究竟需要哪些实验条件，科学家们对现象有了更精确的理解，这为建立一个更精确的理论奠定了基础。

但是，如果实验结果没有被重现，那么心理学的研究成果如何应用？如果科学家们没有在所有的细节上达成一致，知识和理论并不完全站得住脚，那么如何证明这些结果的应用是合理的呢？这种对心理学发现的担心是很常见的，因为人们没有意识到在其他科学中，结果和理论经常在完全确立之前就开始应用了。当然，第2章中已经清楚地阐述过所有的科学理论都有可能被修订。如果我们在应用科学研究结果之前必须确定知识是完全正确的，那么应用就不会发生了。所有领域的应用型科学家尽最大努力使用最准确的信息，同时也会意识到这些信息有可能是错误的。

许多非科学人士会认为医学比心理学要科学得多。但是在实践中药物的不确定性不亚于心理学。例如，在医学上一些重要的与治疗相关的发现，其可重复性的概率还没有心理学中的高（Lehrer, 2010），诊断往往只取决于医生，而非疾病本身（Welch, Schwartz, & Woloshin, 2012），并且新技术经常会导致过度医疗且不会提高治愈率（Saul, 2010）。心理学的知识总是具有一定的概率和不确定性——但在大多数的生物社会科学领域，情况也差不多。

正确看待“真实生活”和“大二学生”问题

本章提到了几个焦点问题，此外有一点很重要，就是我们应该清楚什么是我们说过的，什么是我们没说的。我们证明了对心理学研究的频繁抱怨源自一个基本的误解，不仅针对心理学，而且针对涉及所有科学的一个基本原则。人工条件并不会让实验研究减色，它们只是被创造出来用以分离变量。

我们也已理解了为何人们质疑心理学家不在所有研究中都使用随机样本，并且解释了这种担心是多余的。最后，我们看到，大二学生问题本来是一种合理的关注，但它有时被夸大了，尤其是当人们对心

理学研究的广泛性和多样性不太熟悉时。

尽管如此，心理学家应当始终注意他们的实验结论不要太过依赖于某一种方法或某一特殊被试群体。这一点将在下章讨论。事实上，心理学的一些领域确实被大二学生问题折磨得够呛（Jaffe, 2005）。作为大二学生问题的一剂良药，跨文化心理学仍然是一个亟待发展的领域。然而，研究型心理学家对于自我批评的高度重视，给了我们一个对此持乐观态度的理由（见第12章；Baumeister, Vohs, & Funder, 2009; Lilienfeld, 2011, 2012; Mischel, 2008; Peterson, 2009; Rozin, 2006, 2007, 2009; Simmons, Nelson, & Simonsohn, 2011）。每年各类科学杂志上，都会有文章提醒心理学者注意其方法上的漏洞，或是指出大二学生问题。后者在心理学中是一个受到广泛关注的问题，尚未意识到这一点的心理学家寥寥。因此，尽管我们不应忽视这一问题，同时也应正确看待它。

小结

一些心理学研究属于应用型研究，它们的目标是把研究结果直接应用于特定情境。在这样的应用研究中，研究的目的是要将结果直接推广到自然情境中，样本的随机化和条件的代表性就显得尤为重要，因为研究结果将会直接得到应用。然而，大多数心理学研究不属于这种类型，而是属于基础研究，用以验证有关行为潜在机制的理论。在大部分基础研究中，研究结果通过理论上的修正得到间接应用，从理论产生到应用于某些实践性问题需要一段时间。在这种类型的基础研究中，被试的随机取样和情境的代表性不是关键问题，因为这类研究的重点在于验证理论的普遍性。实际上，在用于验证理论的基础研究中，人为的环境条件是有意创设的，因为（正如第6章所描述的）这有助于把研究的关键变量从所要控制的无关变量中分离出来。因此，心理学实验“不像是真实的生活”这个事实其实是一种优势，而非缺点。

Chapter 8
避免爱因斯坦综合征：聚合性证据的重要性

“生物学实验揭开生命的奥秘！”“思维控制上的新突破！”“加利福尼亚科学家发现了延缓死亡的方法！”如你所见，想仿制一条充斥于媒体头版头条的“突破性”新闻简直就是易如反掌。由于部分缺乏责任感的媒体总是定期炮制这类“头版头条”，难怪大多数科学家都建议公众要以怀疑的态度来对待此类新闻。但是，本章的目的不仅仅在于反对夸大事实、以讹传讹的做法，也不仅仅是提醒人们在评估科学进展报告时必须审慎地考察其来源，我们还想提出一种比前面章节中提到的理念都更为综合、全面的科学进步观。为此，我们将会详细阐述曾在第1章中介绍过的系统实证主义和公共知识。

媒体上这类所谓的“突破性”头条新闻，在很大程度上误导了公众对于心理学和其他科学的认识。一个特别典型的误解就是，它们让公众以为某一科学研究领域中的所有问题都能通过某个关键实验得到解决，或者是某一个重要的灵感成就了理论的进步，并彻底颠覆了先前众多研究者累积的全部知识。这种科学进步观非常符合新闻媒体和网络炒作的胃口，在媒体的运作方式里，对历史的追溯就是呈现支离破碎、缺乏连贯的小型事件。对于好莱坞娱乐业来说，这也不失为一种颇为便利的模式，在那里，事件必须有一个开头和圆满的结尾，含糊的东西都被理得清清楚楚。然而，这只是对科学进步的一种歪曲。如果对此信以为真，就会导致关于科学进步的错误观念，并削弱人们在某一问题上评估科学知识的能力。在本章中，我们将会讨论科学的两个原则——关联性原则和聚合性证据原则，用这些原则描述科学发展，将比“跃进模式”更为准确。

关联性原则

在否定所有科学进步的“飞跃”或者关键实验模式的有效性的同时，我们不是说这种关键实验和理论发展模式从未发生过，相反，科学历史上一些著名案例表明这种模式的确出现过。爱因斯坦提出“相对论”就是迄今为止最著名的一个例子，至此，一系列非凡的理论灵感重新定义了时间、空间和物质等基本概念。

然而，爱因斯坦的成就如丰碑般矗立，让这种科学发展模式统治了公众的内心。这种统治是持久的，因为它与媒体报道大部分新闻事件时所采用的隐含“脚本”高度吻合。人类历史上，像相对论那样遭受了那么多的胡言乱语并不多见（不，爱因斯坦没有证明“一切都是相对的”）。当然，我们的目的不是去批驳这些谬论，而是为了给后面讨论和评估心理学中的理论做铺垫。

在爱因斯坦的理论中，那些被重新定义的关于物理世界的概念是如此地基础，以至于那些通俗读物经常将其等同于艺术领域里的概念变化（一个二流诗人经过重新评估，摇身一变成了天才；一个艺术流派被断言灭亡）。这种做法忽视了概念变化在艺术和科学中最根本的差别。

科学中的概念变化遵从关联性原则，而这一原则在艺术中并不存在，或至少说是极为罕见的（见Bronowski, 1977; Haack, 2007）。就是说，一个新的科学理论，必须与先前已确立的实证事实建立关联。新的科学理论不仅仅要解释新的事实，还要兼容旧的事实，这样才会被认为是一个真正的理论进步。新的理论可以以一种迥然不同的方式来解释旧的证据，但是它必须能解释得通。这些要求保证了科学在原有的基础上持续进步。除非理论解释效力的范围被拓宽了，否则真正的进步是不会发生的。如果一个新的理论解释了一些新的现象，但是没有解释大部分旧的事实，那它将不会被认为是对于旧的理论的超越，因此不会立即取代那些旧理论。

无论爱因斯坦理论中的那些新概念是多么令人震惊（钟表变慢、质量会随速度增加，等等），但它们都遵从关联性原则。在宣告牛顿力学的滞后性的同时，爱因斯坦的理论没有否定那些以牛顿观点为基

础的运动事实，或者是认定其毫无意义。相反，在速度较低的情况下，这两种理论都做出了本质上相同的预测。爱因斯坦理论的高明之处在于，它能够解释更为广泛的新现象（有时是令人吃惊的），而这些是牛顿力学所做不到的。因此，即使是爱因斯坦理论这个在科学历史上最惊人的、基础性的概念重构，也依旧遵循着关联性原则。

消费者规则：警惕关联性原则的无效性

科学发展的“跃进式”模式——我们可以称之为爱因斯坦综合征——让我们误入歧途，以为新的发现必定违反关联性原则。这一观念很危险，因为如果舍弃关联性原则，最大的受益者将是那些伪科学和伪理论的贩卖者。这些理论之所以受到青睐和关注，就是因为它们总被说成是“全新的”。“毕竟，相对论在它所在的时代是新生事物，对吧？”这句话经常被用作一种说辞，以证明某种新鲜玩意儿是正确的。当然，在这个伪科学家虎视眈眈的领域里，先前积累的事实数据看上去似乎是个巨大的绊脚石。然而，事实上，这块绊脚石也无法阻挡这些伪科学家，这是因为他们有两种强有力的伎俩来化解这一麻烦：第一种伎俩我们之前已经讨论过（见第2章），就是解释数据前先将这个理论变得不可证伪，这样就令先前的数据毫无用处了；第二种伎俩是宣称先前的数据与他们的主题无关，因而不予考虑。为了实现“不予考虑”的结果，他们通常强调新理论呈现出“前所未有”的新颖性。

类似“关于现实的全新观念”和“前所未有”这样的语句被频频使用。但实际上，真正的花招还在后面。“新理论”注定如此具有突破性，以至于源于其他理论测试的实验证据都被宣称是与之不相关的。只有能被新理论的框架所兼容的数据才会被考虑，也就是说，关联性原则被完全破坏了。显然，这个理论是如此之新，以至于他们可以理直气壮地说：与之关联的实证证据尚不存在。

如此这般，你就拥有了一个适宜伪科学发展的优质土壤：旧的、“不相关”的数据灰飞烟灭，新的相关数据尚不存在。这种伎俩很容易得逞，因为爱因斯坦综合征蒙蔽了关联性原则。而颇具讽刺意味的是，关联性原则的重要性就是由爱因斯坦理论本身所论证的。

哲学家迈克尔·鲁斯（Michael Ruse, 1999）讲述了一个例子来描述达尔文如何使用关联性原则，并舍弃了某个与其他学科之间缺乏必要关联性的新理论。当时达尔文想探寻一种能与他的自然选择理论相匹配的遗传机制，为此他试图去建立一个所谓“泛生论”的理论。“身体

的各部位都会产生一些小的胚芽，这样胚芽在体内循环，并在性器官处聚集，从而传给下一代。”（p. 64）一个问题是，这个理论和细胞学说并不一致。第二个问题是达尔文没有解释这些胚芽是怎样被运送的，因为输血试验已经证明胚芽不能通过血液传输。基于这两点以及其他一些原因，泛生论在科学阵营里被淘汰出局——“因为它与生物学的其他领域不相兼容”（p. 64）。

同样的情况也适用于心理学，一个新的理论如果否定经典条件反射和操作性条件反射，那么它将无法在心理学中立足，因为它无法兼容行为科学中的其他知识。回忆第6章中对“辅助沟通疗法”的讨论，它之所以不能“治疗”自闭症语言障碍，是因为它打破了关联性原则——如果治疗有效，它将会要求我们重建神经病学、遗传学和认知心理学领域内的知识。这一假设性的疗法与科学中的其他知识没有任何关联。

有这样一个来自心理学的例子。假设有两种疗法被开发出来，用于帮助有严重阅读困难的孩子解决其问题。两种疗法都没有经过实证性的检验：第一种，疗法A是一个训练程序，目的是在音位水平上促进儿童对语言片段的认知；第二种，疗法B通过让孩子蒙上眼睛走平衡木，以训练前庭器官的感受性。

疗法A和疗法B在一个方面上是一致的 ——它们的效果都没有经过直接的实证检验，二者反响都不好。然而，其中一种疗法在关联性原则方面是占据优势的。疗法A与研究文献中的广泛共识具有一致性，在这些研究文献中提到，具有阅读困难的孩子受到阻碍，是因为孩子还没有发展出足够的对于语言片段结构的认知（Snowling & Hulme, 2005; Wagner & Kantor, 2010）。疗法B没有和任何相应的学术共识发生关联。这种关联性的差异预示疗法A是一个更好的选择，即使二者都还没有经过直接的检验。

“跃进”模式与渐进整合模式的比较

把爱因斯坦式的革新视为科学典型的倾向，诱使我们误以为所有科学进步靠的都是重大飞跃。问题就在于，人们倾向于将这些例子泛化成一种观念，认为科学进步理所应当是这样产生的。事实上，很多科学领域的进步靠的都不是某一个突然的突破，而是由一系列构不成重大影响的停顿及前进之间的反复组成的。

科学工作的不确定性是大部分公众所意识不到的。科学实验很少

能完全确定某个问题或支持某一理论，从而排除其他理论。新的理论也很少能够全面超越所有先前存在的相互竞争的概念体系。很多问题的确定并不像科学电影里所描绘的那样，由一个关键实验所决定，而是要等到科学界逐渐开始有了共识，认为支持某种理论的证据比支持其他任何理论的证据要有力得多。

科学家所评估的证据不是来自某个设计得异常完美的实验的数据，与之相反，科学家往往需要去评估来自几十篇实验论文的数据，这些实验各有瑕疵，但每个实验都能提供部分答案。科学发展的这种渐进模式受到阻碍，正是因为爱因斯坦综合征在公众中造成了一种思维定势，认为所有科学都和物理学一样，因为对于物理学来说，科学进步的跃进模式或许是最适用的。

想想遗传学和分子生物学在过去一个世纪中的突飞猛进。这些进步的产生不是因为一个爱因斯坦式的伟人在关键时刻现身，然后搞定了一切。而是，数百个存在瑕疵的实验所产生出来的数十种灵感和洞见，促成了现代生物学的整合。这些进步的发生，凭借的不是革命性地重构一些重大概念，而是几种能站得住脚的不同解释进行长期与反复的交锋和对峙。经过十几年没有定论的实验、无数次的理论构思、争辩与批判，科学家们终于弄明白了基因到底是由蛋白质还是由核酸组成的。但不是通过一次跃进式的改变，他们就达成了新的共识。原子核的发现者恩斯特·卢瑟福（Ernest Rutherford）强调了关联性原则的重要性：“科学家不能仅仅依靠一个人的观点，而是要依靠千万人的智慧。”（Holton & Roller, 1958,p. 166）

卢瑟福的观点强调了另一种区分科学与伪科学的方法。科学总是遵循关联性原则，其特点在于众多个体的参与，而对这些个体的贡献进行评判的标准，是看它在多大程度上加深了我们对自然界的了解。没有哪个单独的个体能够依靠其特殊地位来主导讨论。当然，在第1章，我们已经讨论过科学的这种公共性，相比之下，伪科学经常认为特定的权威和研究者才有接近真理的“特殊”机会。

我们曾提出过两个理念，能为理解心理学的规则提供一个有用的情境：首先，科学上没有哪个实验是被设计得完美无缺的，对任何一个实验数据的解释都存在着不确定性，科学家们评估一个理论，往往不是坐等一个完美的或者关键的实验出现，而是对大量实验的总体趋势进行评估；其次，许多科学在即使没有爱因斯坦的情况下也取得了进步。这些进步是蹒跚而曲折的，而不是通过伟大的“爱因斯坦式整合”那样的阶梯式跃进。

和心理学一样，其他许多科学也都是由那些原本缺乏共同主题的

知识不断积累聚合性证据，在瑕疵中进步。先前的讨论引出了一个证据评估的原则，这个原则在心理学中至关重要。它常被称作聚合性证据原则（或者操作聚合原则）。科学家和那些科学知识的运用者常常不得不作出判断：海量的证据到底说明了什么。在这种情况下，聚合性证据原则就成了一个非常重要的工具。聚合性证据原则对于科学信息的外行使用者来说也是个有效的工具，尤其是在他们要对心理学主张作出评估的时候。尽管对于聚合性证据这个概念所做的详尽性技术讨论将很快让我们晕头转向，但事实上，此概念在实际应用方面的作用很容易理解。我们将通过探索两种方式来表述这个原则，一种是按照“有局限的实验”的逻辑，另一种是按照理论检验。

从极端上讲，导致一个实验出错的方式有无数种（或用术语来说，就是变得混淆）。然而，在某个特定领域中拥有丰富经验的科学家，往往很清楚什么是最容易混淆的因素。因此，当对某一研究结果进行审查时，科学家总能察觉实验中的关键瑕疵。接下来，聚合性证据原则提示我们去审查相关研究文献所呈现的瑕疵模式，因为这类模式要么支持、要么否定我们想要做出的结论。

假设来自大量不同实验的结果都一致支持了某一特定结论。假如实验本身并不完善，我们应该继续去评估这些局限性研究的性质和程度。如果所有的实验都是以同样一种方式出现瑕疵，这些情况将会降低我们对实验结论的信心，因为结论的一致性也许仅仅源于一个特定的瑕疵，而这个瑕疵是所有实验共有的；另一方面，如果所有实验都呈现出不同的局限性，我们对结论的信心就会大增，因为结果的一致性看似并非源自某一个让所有实验结果都混淆不清的干扰性因素。正如安德森（Anderson, 1996）所言：“不同的方法很有可能涉及不同的假设，当一个假设能够通过众多基于不同假设的证伪检验时，我们可以说是得到了一个强有力的结论。”（p. 742）

每一个实验都有助于纠正其他实验在设计方面的错误，只要大量的实验能够得到近似的结果，那么我们就可以说我们的实验证据实现聚合了。即使没有一个实验被设计得十全十美，我们还是得到了一个相当有说服力的结果。因此，聚合性证据原则允许我们将结论建立在大量有些许差异的实验来源之上。这个原则之所以能让我们得出有说服力的结果，是因为这种方法所获得的结果的一致性不大可能是由某个实验程序的特殊性所造成的。

聚合性证据原则同样能以理论检验的形式加以表述。当一系列实验始终支持某个假定的理论，同时又能共同排除那些最主要的竞争性理论时，研究就具有高度的聚合性。尽管没有一个单一的实验能排除所有的可能解释，然而一系列具有一定诊断效力的实验（如果所有数

据都呈现某种特定趋势的话）就能产生一个极具说服力的结论。情况如表8—1所示：

表8-1

	理论 A	理论 B	理论 C	理论 D	理论 E
实验 1	否定	否定	支持	未验证	未验证
实验 2	未验证	未验证	支持	否定	否定

	理论 A	理论 B	理论 C	理论 D	理论 E
总体结论	否定	否定	支持	否定	否定

例如，假设针对某一现象，有五种不同的理论（称它们为A、B、C、D和E）同时存在，且都经过了一系列的实验验证。假设一部分实验以很强的效力检验了理论A、B和C，结果数据否定了A和B，支持了C。再想象一下，另外一些实验则以同样的效力检验了理论C、D和E，结果数据否定了D和E, 支持了C。这种情况下，对于理论C我们就有了强有力的聚合性证据。我们不仅有了支持理论C的数据，还拥有了对抗其他竞争性解释的数据。强调一下，没有一个实验能够检验所有的理论，但是汇总起来，一系列实验就能作出有力的推断。

相反，如果所有已知的研究都只强有力地检验了B、C和E, 并且数据结果支持C并否定了B和E, 那么理论C的说服力就不如前面表8—1中的例子那么强了。原因在于，尽管产生了支持理论C的数据，仍然没有强有力的证据能够排除其他可能的理论（A和D）。情况会像表8—2所示：

表8-2

	理论 A	理论 B	理论 C	理论 D	理论 E
实验 1	未验证	否定	支持	未验证	否定
实验 2	未验证	否定	支持	未验证	否定

	理论 A	理论 B	理论 C	理论 D	理论 E
总体结论	未验证	否定	支持	否定	否定

因此，当一系列实验始终支持某个假定的理论，同时又能共同排除那些非常重要的竞争性理论时，研究就具备了高度的聚合性。尽管没有一个单独的实验能够排除其他可能的解释，但如果将一系列具有部分诊断性的研究按照第一个例子中的方式加以汇总，就能得到一个比较有说服力的结论。

最后，聚合性证据原则能够让我们摒弃一个误区，这个误区的形成是由于我们在第2章对于证伪性的讨论过分简单化所造成的。当时的讨论似乎让人觉得，当第一个与自己的理论相抵触的证据出现时，

这个理论就算是被证伪了。然而事实并非如此。正如理论是被聚合性证据所支持一样，它也要被聚合性的研究结果所否定。

心理学中的聚合性证据

强调聚合的重要性的原因在于，心理学结论往往是建立在聚合性证据原则之上的。这个事实当然并不独特或罕见（在其他很多的科学中，结论也不是基于单一的、决定性的实验证据，而是基于众多结果不甚明晰的实验）。但这种情况在心理学中尤为突出，心理学实验的诊断性往往较低。也就是说，支持某一个理论的数据经常只能排除一小部分可能的解释，还遗留了许多有可能取代这种理论的“候补”理论。其结果是，只有收集并比较来自大量研究的数据之后，才能得到有说服力的结论。

心理学实验具有高度模糊性的因素，这毫不奇怪，因为其研究的问题涉及复杂的人类行为。如果心理学家坦然承认这个事实，然后耐心地去解释这个事实所带来的结果，那么公众就能够更好地理解这门科学。心理学家应该承认，尽管心理科学已然存在并且在不断进步，但这种进步是缓慢的，并且许多结论往往都来自于令人感到折磨的长时间的统合和争论之中。对于媒体经常宣称的所谓突破性进展，我们要永远带着怀疑的态度，但心理学主张所经受的怀疑却是千真万确的。

在心理学中，我们必须像走钢索一样谨小慎微。例如，我们必须抵御这样的诱惑：当证据还不确凿时，就把某一假说当作已经证实了的理论来对待。本书连续几章都反复强调了这种怀疑态度。要注意，不要从相关中推论因果，拒绝接受见证叙述式的证据。与此同时，我们不应该对知识的不完整和最后结论有待探索等事情反应过度，并开始怀疑心理学究竟能否产生有说服力的结论。我们也不应该被“心理学不可能成为一门科学”这种非理性的主张所诱惑。根据这一立场，聚合性证据的原则可用来平衡对假设性知识所作的过度诠释。尽管所有的心理学研究都存在这样那样的瑕疵，但聚合性能让我们获得有说服力的结论。

证明聚合性证据原则的最好方法就是检验心理学中一些仍存在争议的领域。让我们通过一个例子看看聚合性证据原则的重要性。这个问题是，接触暴力电视节目是否会增加孩子们的攻击性行为。对于这个问题，目前的科学共识是：观看暴力电视、电影或录像的确能增加儿童的攻击性行为。这种影响不是很大，但真实存在。科学家们对于

这个结论的信心并非来自某个单一的、权威的研究，而是来自许多研究结果的汇总（Anderson & Huesmann, 2005; Carnagey, Anderson, & Bartholow, 2007; Feshbach & Tangney, 2008; Fischer, Greitemeyer, Kastenmüller, Vorgrincic, & Sauer, 2011b）。这一研究结论适用于电视游戏、电视及电影（Carnagey et al., 2007; Sheese & Graziano, 2005）。这些研究所采用的研究设计、被试规模以及特定技术都有很大差别，但现在能够清楚发现，这些差别是此领域内各种研究的优势，而不是弱点。

尽管电视所属的产业能给孩子带来负面影响，且证据十分确凿，但那些电视网和电视游戏产业的老板们还是自然而然地会抵制这些证据。他们发起了一场误导公众的运动，利用的正是公众“不能意识到研究结果是建立在许多研究的聚合上而非某个单一的、具有决定意义的证明上”这一特性（Seethaler, 2009）。电视网和电子游戏公司不断挑选个案并暗示，只要证明每一个研究都存在瑕疵，就可以全盘否定总体的结论。尽管社会科学研究者也许会去回应对于某个具体研究的批评，但并不能就此认为研究者总是轻易承认某个特定研究存在着缺陷。关键的区别就在于，研究者拒绝这样的暗示，即承认某个特定研究存在瑕疵，就否定了“影视暴力会对攻击性行为产生影响”这一普遍的科学共识。其原因就在于，普遍的结论来源于聚合性。即使是不包含这类瑕疵的研究，其结果也会指向同一方向。这一研究当然也有其自身的问题，但其他研究对此进行修正之后也产生了相似的结论。

例如，关于这个问题，早期研究揭示了观看暴力节目的数量和儿童攻击性行为之间的相关。这些相关证据不能被视为因果关系，指出这一点是非常正确的。也许是第三个变量导致了这种关联，也许更有攻击性的儿童选择去观看更多的暴力节目（方向性问题）。

但是科学团体的结论不单单是建立在相关证据之上的。研究者不仅对两个变量之间的关联进行简单的测量，还使用了更为复杂的相关技术，这些相关技术允许研究者得出一些因果性质的试探性结论（其中一个如偏相关，在第5章中曾提到）。在这类技术中，有一种方法采用纵向设计，即在不同时间点测量相同的两个变量——在这里就是电视暴力和攻击性。由这一设计所得到的相关模式可以告诉我们二者是否有因果联系。有人已经进行过这类研究，得到的结果表明：观看暴力节目的确有可能增加人们日后的攻击行为。

同样，有人批评纵向相关技术尚存在争议，这并非毫无道理，因为它确实有争议。关键在于，“电视暴力和攻击性行为之间存在因果联系”这一结论依靠的并不完全是简单或复杂的相关证据，研究者还进行了无数的实验室研究，在这些研究中，电视暴力的数目得到了直

接操纵，而不仅仅是被评估。在第6章，我们曾讨论过变量的操纵，操纵与随机分配等控制手段共同使用，就能避免相关研究在解释问题时存在的不足。如果有两组儿童，在其他变量均得到了实验平衡之后，仍然表现出不同水平的攻击性行为；如果这两组儿童的唯一的区别就是一组观看暴力节目，而另一组没有观看暴力节目，那我们就能做出正确的推断：被操纵的变量（电视暴力——自变量）导致了结果变量（攻击性行为——因变量）的变化。这个结果在大部分实验研究中都出现了。

这些研究已经激起了一些“这不是真实的生活”的非议，这种非议连同那些毫无根据的指责，在先前的章节中都曾讨论过。无论怎样，电视暴力的影响并非针对某个特定群体的儿童，因为这些结果在美国不同的地区和世界的不同国家都得到了印证。使用不同的实验情境、不同电视节目作为实验刺激的各个研究都得到高度一致的结果。

重要的是，从现场实验而非实验室实验中也得到了相同的结论。一种叫作现场实验的设计也被用来研究电视暴力/攻击性行为问题。这类研究设计的存在提醒我们，不要认为实验情境和实验设计之间存在必然的联系。有时候人们认为，我们只有在实验室里才能操纵变量，在非实验室情境中只能进行相关研究。这个想法是不正确的。实验室里也常常进行相关研究，而非实验室情境下也常常可以操纵变量。尽管有时在非实验室的情境下操纵变量进行现场实验（有的在第6章中有所涉及）需要相当大的创造性，但这一方法在心理学领域中被越来越普遍地采用。

例如，一项最近的现场实验检验了犯罪事件的“破窗理论”（Keizer, Lindenberg, & Steg, 2008）。该理论认为，即使是看上去偶然为之的社会失序指标（破损的窗户、涂鸦等），也能传递“该地区犯罪很常见”的信号，并因此提升犯罪率。凯·科泽尔和他的同事在停放自行车的小巷中设计了两个情境。在控制情境下，小巷中有禁止涂鸦的标语，并且小巷中没有涂鸦。在实验情境下，小巷中有禁止涂鸦的标语，但是小巷的墙上有很多涂鸦。两个小巷中的自行车把上都有纸质传单。科泽尔和他的同事发现，在实验组有69%的被试乱扔传单（扔在地上），而控制组只有33%。当然，现场实验本身是有缺陷的，这些缺陷往往是其他研究的强项。总的来说，将观看电视暴力和儿童攻击性行为增加联系起来，所使用的证据并非仅仅依靠某一研究甚至某一类型的研究。

这种情形类似于吸烟和肺癌的关系。吸烟的人死于肺癌的几率比不吸烟的人高出15倍（Gigerenzer et al.，2007）。烟草公司的老板们经常试图去误导公众，暗示吸烟导致肺癌的结论只基于个别研究，

然后便开始批评个别研究（Offit, 2008）。恰恰相反，支撑这个结论的是很多聚合性的证据。来自于不同研究的数据的聚合性是很强的，这些数据的聚合性不会因为对某个研究的批判而彻底改变。

事实上，在这里有必要讨论一个类似肺癌起因的科学问题。医学诊断和治疗中的许多决策，都建立在不同研究结果能否汇聚为一个结论的基础之上。例如，当流行病学调查（可以说是一种涉及人类的现场研究，目的在于寻求某一疾病与环境及地理因素的关联）、精确控制的动物实验以及人类被试的临床实验等不同类型的研究结果，都趋向于汇聚在一个结论上时，医学界才会对这一结论抱有较大的信心，认定这一结论是可靠的，医生们才愿意在这些证据的基础上实施治疗方案。

然而，所有这三种类型的研究都有其各自的缺陷。流行病学研究经常是相关性的，在变量之间存在虚假相关的可能性很高。实验室研究能被高度控制，但实验对象往往是动物而不是人类。医院环境下的临床试验在真正的治疗环境中使用人类作为被试，但仍有很多控制的问题，因为存在安慰剂效应和治疗病人的医疗团队的期望效应。就像吸烟和肺癌的例子中那样，尽管每一种研究都存在问题，但当来自不同方法的数据能较强地汇聚起来的时候，医学研究者们就能够做出有说服力的结论。这与心理学家可以用聚合性证据原则来帮助他们做出电视暴力对攻击性行为有影响的结论一样。

然而，对于公众来说，聚合性证据原则常常是难以理解的。例如，华盛顿大学流行病学专家大卫·迈克斯（Michaels, 2008）讲述了这样一个法庭案例：庄纳诉通用电气案，此案涉及有毒物质的伤害。法官发现每个呈现的科学证据都有缺陷，然后就将所有的证据都弃置一旁。迈克斯提醒我们，“在真实世界中，科学家不会这样做。他们会考虑每个证据的优缺点。尽管实验研究中得到的证据是有缺陷或者局限的，但仍完全有可能得出一个合理的结论。这种事情始终在发生。”（p.163）

有时，聚合性证据原则不为大众所知。另一些时候则是为了达到政治性目的或是经济目的而被有意识地忽视。很显然，烟草公司专家和高管都在试图混淆大众对吸烟导致肺癌的聚合性证据的理解，其实他们知道聚合性原则，但他们希望蒙蔽大众。一个和吸烟/肺癌相关的例子就发生在当下。在科学上有一个很强的聚合证据显示边打电话边开车（或者开车的时候因为电子仪表设备分心）极其危险，并且是导致车祸的重要原因。然而手机公司和汽车公司——就像之前的烟草厂一样——试图蒙蔽大众有关这个结论在科学上高度聚合的事实（Conkle & West, 2008; Institute for Highway Safety, 2005; Kunar,

Carter, Cohen, & Horowitz, 2008; Levy, Pashler, & Boer, 2006; McEvoy et al., 2005; Redelmeier & Tibshirani, 2001; Strayer & Drews, 2007; Strayer & Johnston, 2011）。

科学共识

评估电视暴力影响的问题是一个典型例子，告诉我们：在心理学中，数据最后是如何累积起来用于解决问题的。尤其是在社会急切关注的领域，切记，这些问题的答案只能在大量不同研究结果实现融合之后缓慢地出现。用一个简单原理来总结：在评估心理学的实证证据时，心中要想的是“科学共识”，而不是“重大突破”；是“渐进整合”，而不是“大步飞跃”。

对“共识而不是突破”的不接受阻碍了公众对人类行为会造成全球变暖的理解（Geant, 2011; Jordan, 2007; Nijhuis, 2008）。许多政治团体不喜欢有关人类能源消耗、汽油使用、经济活动中的碳排放会对环境造成消极影响的证据，因为这些证据与他们的政治意图相左。因此，很多政治团体会攻击个人研究，并且利用媒体运动让这种攻击拥有更广泛的群众基础。他们想制造这样一种印象：因为个人研究是值得怀疑的，所以在人类行为与全球变暖的关系上存在巨大的科学争议（Grant, 2011; Michaels, 2008; Nijhuis, 2008）。事实上，争议并不存在，因为结论不是通过单独一个研究得出的。1993到2003年间，有关全球气候变化的论文发表了900多篇，压倒性地汇聚成结论：全球变暖有人类行为的影响（Oreskes, 2004; Oreskes & Conway, 2011）。一项研究并不能决定整体的结论，显然，否定一项研究也不能改变结论。然而，政治团体希望在公众中制造疑虑，而且他们也成功做到了。调查显示，大约50%的民众认为科学家在结论上还存在争议，事实上强聚合性结论已经得出了（Frazier, 2009）。

不幸的是，媒体“公说公有理，婆说婆有理”的取向成为了否认全球变暖的推手，媒体那种“一方面……另一方面”的报道方式暗示在此问题上存在巨大的争议，虽然事实并非如此（Oreskes & Conway, 2011）。近年来，一些媒体开始采取行动，以阻止政治团体剥夺公众对于“科学结论来自聚合性证据和共识”的理解。

2007年8月13日，《新闻周刊》杂志的封面是一个巨大的标题“全球变暖是个骗局”和一个星号。封面的左下方，星号解释了这个标题纯粹是个玩笑！这一主题的封面文章是关于资金充足的政治团体如何说服民众相信“人类造成环境改变是有争议的”这一说法。杂志上

刊登了关于被称为“否认机器”的长篇文章，所谓否认机器，指的是那些以“单一研究是有缺陷的”为由，成功让大众相信关于全球变暖存在疑问的政治团体（Oreskes & Conway, 2011）。文章描述了政治团体是如何利用民众的误解使观众错信这个问题一定是由单一的关键性研究所决定的。

因此，科学作家芭芭拉·坎罗威茨和克劳迪娅·卡尔布（Kantrowitz & Kalb, 2006）警告说，媒体对医学研究的报道有积极的传播意义，但是也可能事与愿违，公众如果没有学习过聚合性原则，真正的理解会更少。他们指出，在医学上，科学是一点一点、一小步一小步地进步的，而不是突破性地进步。但是，媒体总是将事情描绘成相反的形式。

研究方法和聚合性原则

聚合原则同样也意味着，我们应当乐于看到多种不同方法应用于心理学研究的各个领域中。因为不同的研究技术各有其优势和不足，用于获得特定结论的各种方法之间呈现一种相对的平衡是比较理想的。心理学长期以来都因过于依赖基于实验室的实验技术而饱受诟病。批评的效度取决于作为讨论焦点的特定研究领域。然而，一种确定无疑的趋势是，近年来，心理学各个领域都已经开始使用不同的研究方法了。例如，由于过度依赖实验室技术，社会心理学家遭受的批评可能是最多的，但社会心理学家已经开始转向了更富想象力的现场设计，以寻求聚合性的证据来支持他们的理论。

比如说，有大量的关于被称作旁观者效应的研究。旁观者效应是指，一些人在看到他人处于危难之中时并不施以援手（Fischer et al., 2011a）。当有更多潜在的帮助者在场时，帮助行为出现的可能性会降低。早期研究这种现象的研究者清楚地知道，这些仅凭被试在实验室里的反应而做出的结论太过单薄了。因为在实验室中，被试都是在自愿报名到实验室来参加实验之后才目睹紧急事件的。为此，拉坦和达利设计了另外一个著名的实验，希望在另一个情境中重现这一现象。他们找到一个愿意合作的卖酒的商店，该商店同意假装店里发生了盗窃事件。当收银员在店铺的后面为一个“顾客”拿啤酒时，该“顾客”（实际上是研究者的同伴）拿起一箱啤酒走出店门。这一幕总发生在收银台前一个或两个真正的顾客的眼皮底下。收银员回来后问这一个或两个顾客：“嗨，刚才在这里的那个人到哪儿去了？你看见他离开了吗？”这样，就给了顾客一个机会向收银员报告刚才发生的盗窃事件。与实验室实验的结果吻合：当旁观者在场的时候，向收银员

报告盗窃案的行为受到了抑制。

在第10章中，我们将讨论许多带有概率性质的决策原则，这些决策原则最早都产生于实验室，但都经过了现场式的检验。例如，研究者使用源于实验室的原理解释了理疗师、股票经纪人、陪审员、经济学家及赌徒在各自所处情境下是以何种方式做出概率推理的（Adler, 2009; Hilton, 2003; Kahneman, 2011; Stanovich, 2011; Thaler & Sunstein, 2008; Zweig, 2008）。

实验与非实验结果的聚合性也成为教育心理学领域的突出特点。例如，针对不同课程安排所做的实验研究和现场研究都表明，早期语音教学有助于阅读技巧的习得（Ehri, Nunes, Stahl, & Willows, 2001; Pressley, 2005; Snowling & Hulme, 2005; Vellutino, Fletcher, Snowling, & Scanlon, 2004）。

同样也应记住，聚合性研究在支持初始假设方面并不总是正向的。有时，聚合性研究得到负向结论——初始的假设不能得到证实。在教育心理学领域，在研究学习方式方面有一个这样的案例。长期以来，人们相信存在一种方法能让教师测量出每个孩子的“学习方式”。我在这里不会论及特定的方式，因为不同的作者对“方式”包含什么有不同的理解（这有时候也是问题的一部分）。不管怎样，教师们随后理应“教授”这些方式——能够使孩子们的成绩变好（有时也称如果采用这种方式，所有的学生都能同等地获得成绩上的大幅提升）。问题是，对这个想法进行数以百计的研究之后，研究没有证实这种观点（Lilienfeld et al., 2010; Pashler, McDaniel, Rohrer, & Bjork, 2009; Stahl & Kuhn, 1995），也没有可重复的证据显示，教师可以将教学和这些方式“匹配”以达到更好的学习效果。

向更有效的研究方法迈进

对于某个特定问题的研究，通常是从相对较弱的方法过渡到可以做出较强结论的方法。例如，研究者对某个特定假设的兴趣，常常源于某个异常感兴趣的特殊个案。正如我们在第4章中讨论的，这就是个案研究的真正作用：为更有效力的进一步研究提供一些假设，同时激发科学家们用更为严格的方法去研究这些假设。个案研究之后，研究者多采用相关研究来确认变量之间是否存在真正的关联，而不仅是存在于几个个案中的巧合现象。如果相关研究证实了变量之间的关联，研究者就开始尝试采用实验法来对相关变量进行操纵，借以找到变量之间可能存在的因果关系。这个递进的顺序就是：从个案研究到

相关研究，再到操纵变量。

讨论“向更有效的研究方法迈进”为我们提供了一个纠正错误概念的机会，这个错误概念源于第5章的讨论，那就是“相关研究在科学中没有什么用处”。的确，当一个因果关系的假说需要验证时，操纵变量的研究方法更受青睐。然而，这并不意味着相关研究对于知识的获得毫无帮助（West, 2009）。首先，许多科学假设是以相关或者不相关的形式来表述的，因此这类研究是在直接验证这些假设；其次，尽管相关并不意味着因果关系，但因果关系一定包含相关。也就是说，如果一个相关研究不能肯定地证实因果关系的假设，那它可以起到排除这一因果假设的作用；最后，相关研究或许比它们看上去更有用，因为最近新发展的复杂相关设计可以让研究者做出有限的因果推论。我们在第5章讨论了偏相关这种复杂的相关技术，这一技术有可能检验出变量间的关联是否能够被第三变量所解释。

然而，最重要的原因可能在于有时出于道德或伦理的考虑，我们无法对一些变量进行操纵（例如，营养不良或肢体残障）。而另外一些变量，诸如出生顺序、性别、年龄等，则因其无法被操纵而具有天然的相关性，涉及它们的科学知识也因此必须建立在相关证据基础上。当然，这一情况并不是心理学领域所独有。天文学家们显然无法操纵所有影响其研究对象的变量，然而他们依然能够做出结论。

在健康心理学中，有一个研究方法演进的例子，它涉及A型行为模式和心脏病之间的关系（Chida & Hamer, 2008; Martin et al., 2011; Matthews, 2005; Suls & Bunde, 2005）。最初，A型行为模式这一概念源于两位心脏病专家的观察，这两位医生从他们一些病人的行为中发现了一种稳定的模式，这种行为模式包括时间紧迫感、飘忽不定的敌意，以及对成就的极度渴求。于是，一些医生通过对少数个案的观察，提出了“A型人格”这一想法。这些个案研究提出了这个概念，但并不能作为有力证据来证明这种特定的行为模式是导致心脏病的原因之一。要证明这一点，需要的不仅是少数几个个案研究，它还需要由心脏病专家、生物化学家和心理学家团队数十年的努力。

很快，这个研究从永远也不可能证实假设的纯粹个案研究，转向了更有效力的研究方法。研究者发展和检验了A型行为模式的操作性定义。大范围的流行病学研究证实了A型行为和心脏病之间的相关性。然后这种相关研究工作就变得很复杂了。研究者使用复杂的相关技术来搜寻潜在的第三变量。由于行为模式与其他传统心脏病风险因素中的一种（例如吸烟、肥胖和血液中胆固醇水平）存在相关，因此A型行为和心脏病之间有可能存在虚假相关。当其他的变量在统计上被排除后，A型行为模式和心脏病之间仍然具有关联。最后，研究者

采用了实验研究对变量进行操纵，以期证实二者间是否具有因果关系。一些研究试图去验证是否某些生理机制影响了两者之间的关系，并以动物作为被试——某些人所谓的“不是真实的生活”的研究方法。另外一些研究则以犯过心脏病的人为被试。这些被试被随机分配到两个组中的一组。一个组接受咨询，帮助他们避免传统的风险行为，例如吸烟或者吃高脂肪食物；另一组在接受同样的咨询的同时，还接受了一个以减少他们的A型行为为目的的训练项目。三年之后，在接受A型行为辅导的病人中，心脏病复发的情况要明显少很多。

简而言之，证据汇聚起来支持了“A型行为模式是导致心脏病的重要原因”这一假设。对这个问题的研究提供了一个很好的范例，从中我们能清楚地看到，研究是怎样从一个感兴趣的个案研究走向相关技术，最后到可以操纵变量的实验研究的。

我们能从这个例子中得到的最后一点经验就是，科学概念总是在不断地演进。这个论点是在第3章讨论操作性定义时首次提出的。最近的研究似乎表明，将A型行为与心脏病之间的关系说成是整体性的显得过于简单化了。原因在于，只有该概念中的特定成分（特别是对抗性敌意）才与心脏病有关联（Chida & Hamer, 2008; Matthews, 2005; Suls & Bunde, 2005）。因此，这是个很好的例证，从中可以看出，随着科学的进步，它是如何不断地揭示特定的关联，以及理论概念是如何被细化的。

不要对矛盾数据感到绝望

聚合性原则的最后一个启示是，当一个问题的最初研究结果看上去有些矛盾时，我们不应当对此感到绝望。在科学中，证据融合的过程就像投影仪慢慢将一张未知的幻灯片的焦点调清晰。起初，屏幕上的模糊影像可能代表任何东西。接着，随着一点点地调整焦距，虽然这个图像仍不能被清楚地识别出来，但许多其他的可能假设也许会被排除。最后，当焦距调准，就可以非常有信心地作出最终的判断。证据融合过程就好比一个调焦过程。幻灯片的模糊影像就如同互相矛盾的数据，或者是那些支持多重假设的证据。

因此，研究早期所获得的矛盾数据不应该让我们对发现真相感到绝望。类似的情况不光发生在心理学领域，同样也发生在一些相对成熟的科学中。的确，公众经常意识不到科学中经常会得到一些矛盾的数据。这些矛盾只不过是因为我们对问题理解得还不够充分，这些矛盾还可能仅仅是偶然事件（我们将会在第11章中对此展开讨论），或者源于不同实验在方法上的细微差异。在达成共识之前，其他许多科学也都经历了令人困扰的不确定时期（Ioannidis, 2004; Lehrer, 2010; Simonton, 2004）。这种情况在医学领域一直在上演。例如，研究证实每天服用一小片阿司匹林可以预防心血管疾病。然而，关于阿司匹林预防癌症的研究结论是极为混乱、不确定和不聚合的。阿司匹林通过抑制环氧酶（COX）来消除炎症。因为环氧酶也和肿瘤形成有关，所以人们认为每日服用阿司匹林可能会起抑制作用。但是根据这种推测而实施的真实研究获得的结果却并不一致。一些研究者认为，这种分歧的结果是由于最佳剂量水平还未被找到。无论这个问题最终如何解决，都说明不确定性总是先于科学问题的解决。美国癌症协会迈克尔·图恩（Michael Thun）博士承认，当公众无法理解科学的机制，以及结论总是从缓慢发展的聚合性中逐渐产生时，他们会感到沮丧（Associated Press, 2007）。他承认，对于公众来说，理解“为什么我们不能找到像阿司匹林这种非常常用的药的所有效果”确实非常困难。然而这个例子刚好证明，得出一个因果性结论是多么不容易。就像我们在本书中看到的那样，得出一个因果性的结论并不容易。得出一个确定结论之前会经历长时间的不确定性，这种情况也并非心理学所独有。

作家马尔科姆·兰德威尔（Malcolm Gladwell, 2004）讨论了人们为何很难理解医生对于乳腺X光片的作用还存在着分歧。这是因为乳腺X光透视在大多数人看来是如此“精确有力”，以至于他们认为仅凭它就能作出确诊。其实这些人不理解，医生的诊断虽必不可少，但乳腺X光片评估和疾病预测从本质上来说是具有概率性的（Gigerenzer et al., 2007）。格兰德威尔说：“图片保证确定性，但它不能兑现这种承诺。经过40年的研究之后，对于女性在50岁至69岁期间接受乳腺X光透视的益处，仍然存在着不小的分歧。进一步的争议则在于，是否有足够的证据能够证明，50岁以下和70岁以上的女性定期需要接受乳腺X光透视检查。”（p. 81）然而，格兰德威尔继续谈到，和心理学领域一样，在医学领域里，知识即使不确定也依然有用：“答案是乳腺X光透视不需要完全准确无误才能拯救生命……它没有我们想的那么好。但总归比没有它要强”（p. 81）。

在心理学和其他科学里，将来自不同研究的证据整合起来形成一个结论，已经能够通过一种更为正式的方法来实现，这就是一种叫做元分析的统计技术（Borenstein, Hedges, Higgins, & Rothstein, 2009; Card, 2011）。在元分析中，具有相同研究假设的若干研究结果以统计方法结合起来。两个实验组的比较得出的效应可以纳入一个常规的统计矩阵中，这个矩阵能进行研究之间的比较。接着，这些结果以一种标准化的方式加以统计整合。如果整合过程达到了一定的统计学标准，就能形成一个关于这些效应的结论。当然，在某些情况下，有可能无法确定地得出一个结论，这时元分析的结果就是非结论性的。

越来越多的评论者开始呼吁，应更加重视元分析，并将之视为一种消除科学领域内相互对立研究之间的不断争议的方法。这种方法有助于终止这种“公说公有理，婆说婆有理”的争论。对元分析的强调也揭示了一种观点：专业杂志上常见的观点对立可能只是表面现象，实际上我们拥有更多可靠和有用的发现。

国家阅读评审小组（NRP, 2000; Ehri et al., 2001）对一些关于阅读教育的研究所做的元分析就证明了这一点。例如，他们得出结论，对38个不同的研究结果的元分析“有力地支持了这一观点，即相比其他课程提供的非系统或非语音教学，系统的语音教学在孩子的成长中发挥了更大的作用”（p. 84）。在报告的另一部分，NPR报告说，对于52个语音意识训练研究的元分析说明，“教孩子掌握在语言中运用声音，能帮助他们学会阅读，在不同的教学、测验及参与者的个性条件下，其效应量都远远大于随机水平，并且虽然这些效应有大有小，但大部分都处于中等水平”（p.5）。

在健康心理学领域也有类似的情况。千田和哈默尔（Chida,

Hamer, 2008）对来自281个有关A型行为模式的敌对性和攻击性与心血管反应（心率和血压）之间关系的研究数据进行元分析，试图建立它们之间真正的关系。在另一个例子中，柯里尔、内梅尔和贝尔曼（Currier, Neimeyer, & Berman, 2008）对丧亲者心理治疗进行干预的61个控制研究得到的数据进行元分析。但他们的元分析结果令人失望，心理疗法的干预在刚失去亲人之后有即时的效果，但在随后的时间里却没有积极的影响。

接下来的元分析的结果提醒我们，元分析的结果并不总是积极的。也就是说，我们不能总是认为只要对一系列的不同研究进行元分析就能得到点什么。我们经常会发现，当我们将许多不同的研究结果合并的时候——什么都没证明！例如，迪特里希和简素（Dietrich, Kanso, 2010）用元分析分析了超过70个有关创造力的神经生理关联性的研究。将若干神经影像的研究结果进行整合后，他们发现发散思维没有特定的神经关联。总的来说，他们发现“发散思维不能由任何单独的心理过程或是脑区决定，也不和右脑、散焦注意、低唤起、alpha波同步化这些经常被假设的方面存在特定相关”（p.822）。我们在这要表明的是，元分析有时会产生消极的结论。

元分析包含几十个研究（有时是几百个），这一事实隐含了这样的信息：每一个单独的研究只呈现了更大量研究中的冰山一角。另外，当我们担心心理学的一些领域进展缓慢的时候，我们应该知道“低产”是医学和其他许多科学研究领域的共同特征。

美国心理学会的一支工作团队在心理学期刊上所做的关于统计方法的一番阐述，为本节内容提供了一个恰当的总结（Wilkinson, 1999）。这个工作团队说：“研究者不应仅针对单个研究的结果作出解释，就好像其他文献所报告的结果与之毫无关系似的。”（p.602）不同研究结果之间达成聚合效应，才有利于推动科学进步。一个研究的结果也只有通过针对特定问题的诸多研究获得聚合性解释，才是有意义的。

小结

在这一章中，我们看到，为何“跃进”模式对于心理学来说是一种糟糕的模式，以及为什么“渐进整合”模式提供了一个更好的框架，凭借这个框架，我们就能够理解心理学中的结论是如何形成的。聚合性证据原则描述了心理学上研究结果是如何被整合的：没有一个实验是可以一锤定音的，但是每一个实验至少都能帮助我们排除一些可能的解释，并让我们在接近真理的道路上向前迈进。多种不同方法的使用让心理学家更为确信，他们的研究结果是建立在稳固的实证基础上的。最后，当概念上的变化发生时，它必须遵循关联性原则：新的理论不仅要能解释新的科学数据，还必须能解释已有的数据。

Chapter 9
打破“神奇子弹”的神话：多重原因的问题

在第8章里，我们聚焦于聚合性操作的重要性，以及寻求一种能够更有效地在变量间建立联系的研究方法。在这一章中，我们将超越两个变量间的单一联系，强调另一个重要的观点，那就是人的行为是由多重原因共同决定的。

任何一个特定行为都不是由某个单独的变量引起，而是由许多不同的变量共同决定的。认定变量A和行为B之间存在显著的因果关系，并不意味着变量A就是引起行为B的唯一因素。例如，有研究者发现，收看电视的时间和学业成绩之间存在相关，但不会就此认为收看电视时间是影响学业成绩的唯一因素。道理很简单，学业成绩在一定程度上还受到大量其他变量的影响（例如，家庭环境、学校教育的质量，等等）。实际上，相对于这些变量，看电视只是影响学业成绩的一个次要因素而已。同样地，收看大量的电视暴力也不是使儿童表现出攻击行为的唯一原因，它只是众多影响因素中的一个。

但人们常常忘记行为是由多重原因决定的，他们似乎要去寻找那颗所谓的“神奇子弹”——即他们感兴趣的、造成行为的唯一原因。心理学家希尔多·瓦茨（Theodore Wachs, 2000）以人们试图解释1998至1999年间发生在美国的校园枪击案的方式作为例子指出，人们认为涉及的原因包括枪支容易获得、父母对孩子较低的关注、互联网信息、影视暴力、同伴影响和精神疾病。瓦茨认为：“很少有人觉得校园枪击案激增是上述原因共同作用的结果，任何解决方案都不应只针对某一个潜在的原因。”（p. x）

和本书中谈到的许多其他原则一样，具备原因多样性的观念非常重要。一方面，它提醒我们不要过于依赖单一的原因解释。因为这个世界盘根错节，影响行为的因素也多样而复杂。虽然我们可以证明某一变量引起了某一行为，但并不代表已经发现了影响该行为的唯一原因，甚至是最重要的原因。为了对某种特定行为作出全面的解释，研究者必须探讨各种不同的变量对它的影响，并把这些研究结果整合起来，才能完整地描绘出所有与该行为有关的因果关系。

另一方面，虽然说某个变量只是影响特定行为的众多因素之一，并且只能解释这一行为的一小部分，但并不是说这个变量就是无足轻重的。首先，这一关系可能具有深远的理论意义。其次，这一关系可能具有应用价值，尤其当这个影响变量是可以进行人为控制的时候，如前面提到的电视暴力的例子。如果控制了这一个变量，能够使每年的暴力事件降低1%，那我想没有人会认为它是无关紧要的。总之，如果问题行为至关重要，那么懂得如何去控制其中一个哪怕非常小的原因也具有非凡的价值。

罗森塔尔（Rosenthal, 1990）举过一个治疗心脏病的例子，在一个实验中，某种治疗方案能将患者存活率提高不到1个百分点；然而，即使这样，这个结果也被认为是意义太过重大，以至于基于伦理考虑，实验者不得不提早终止研究：既然实验治疗结果这么有效，对于那些被随机分配在控制组的病人，让他们仍然使用安慰剂显然是违背伦理的。同样，任何能够将机动车死亡率降低1%的因素都至关重要——每年都能挽救450条生命。将凶杀案案发率降低1%，则每年能挽救超过170条生命。总之，一个结果是由多重变量决定的这一事实，并没有降低任何一个与结果存在因果相关的变量的重要性——即使这一变量仅能让结果产生很小的变化。

交互作用

原因多样性的观点引出了另一个重要概念，那就是交互作用。这个概念在许多方法论的书上都有详细的介绍，因此这里不再赘述，只是稍提一下，一个影响行为的因素可能会由于其他因素的出现或不出现而产生不同的效应。这就是我们常说的交互作用——一个自变量的影响效果依赖于另外一个自变量的不同水平。

来看看这样一个例子：有研究者考察了一组青少年的学业平均成绩，想看看一些生活事件（如转学、青春期发育、早恋行为、迁居和家庭破裂等）是否会对学业产生影响。他们发现上述生活事件加在一起，是导致学业不良的关键因素。单一的因素不能产生巨大的影响，但当几个因素结合在一起时就会产生相当大的影响（Simmons, Burgeson, Carlton-Ford, & Blyth, 1987）。

另一个类似的例子是迈克尔·努特（Michael Rutter, 1979）对儿童精神疾病相关因素方面的研究进行的综述，他提出：

> 第一个引人注目的发现是，在实验中，那些被单独分离出来的慢性压力并未增加精神疾病的风险……这些风险因素单独作用时，没有一项与儿童的精神疾病存在关联：这些儿童患精神疾病的风险也不会比没有家庭压力的儿童高。然而，当任何两种不同来源的压力同时作用时，患病的风险就超过原来的4倍。若是3种或4种压力来源同时作用，那么患病的风险更是增大了好几倍。很明显，这些慢性压力的共同作用远远超过其各自效果的累加，因为几种并发压力之间存在交互作用，才令其总体效应远远大于单个压力效应之和。（p. 295）

当诸如努特所描述的交互作用发生时，要理解其发生的逻辑，可以先想象一个风险量表，得分80~110代表低风险，110~125代表中等风险，125~150则代表高风险。假设我们发现儿童在无压力情况下的平均风险得分为82, 在压力因素A作用下的平均风险得分为84，而在压力因素B作用下的平均风险得分为86。当研究因素A和因素B两者对儿童的共同影响时，如果发现风险指数达到了126, 也就是说，联合的风险指数远远超过了独立研究单一因素时所预测的结果，就说明了

因素A和B之间存在着交互作用。

发展心理学中也有许多类似努特所描述的例子。研究者邦尼·布瑞特米亚（Bonnie Breitmeyer）和克雷格·拉米（Craig Ramey）研究了两组婴儿，一组是非最佳围产期的婴儿，另一组是正常婴儿。在这两组婴儿出生后，再把他们随机分配成两组——实验组及控制组，然后对实验组实施一个特别的育婴方案，该方案是为了防止出现轻微智力迟缓而设计的。控制组的婴儿则没有得到任何特殊的照料。当这些孩子长到4岁的时候，对他们的认知发展能力进行测试，发现在特别育婴方案下，非最佳围产期出生的儿童与正常儿童在认知能力上没有显著差异。但是，没有得到特殊照料的控制组中，那些非最佳围产期儿童的表现低于正常儿童的认知发展水平。该研究中，生理和环境因素的交互作用说明，一个复杂的行为结果（认知发展）是由多种因素决定的。当非最佳围产期出生的儿童得不到适当的照顾时，就会出现负面的认知发展结果。研究者们总结道："这个研究结果支持了这样一个理论架构，即对于那些在社会经济条件较低的家庭中成长的儿童而言，先天的生理缺陷和后天不良的环境因素会成为他们发展中的累积性危害因素。"（Breitmeyer & Ramey, 1986, p. 1151）

很多消极的行为和认知后果都伴随着相似的逻辑——许多生物和环境变量存在交互作用的情景亦在此列。例如，5-HTT基因的变异与人类的严重抑郁有关（Hariri & Holmes, 2006）。具有一种变异型（S等位基因）的人比有另一种变异型（L等位基因）的人更可能罹患严重抑郁。然而，有S等位基因的人只有经历多重生活创伤性事件——例如儿童时期被虐待或被忽视、失业、离婚时，患抑郁症的风险才会较大。这类基因—环境交互作用在发展心理病理学领域非常普遍（Dodge & Rutter, 2011）。

例如，单胺氧化酶A（monoamine oxidase A, MAOA）基因的变异与反社会行为的关系就与之类似。只有当诸如儿童虐待、出生并发症抑或不良家庭环境等其他风险因素出现时，基因的一种变异型才会增加反社会行为出现的概率（Raine, 2008）。最后一个例子是思维反刍和抑郁的关系。思维反刍的倾向能够预测抑郁症状的时长，但它与认知风格存在交互作用——只有伴随着消极认知风格，思维反刍才能够预测抑郁症状的持续时间（NolenHoeksema, Wisco, & Lyubomirsky, 2008）。

积极的结果也可用多种因素及其之间的交互作用来解释。奈特等（Knight et al., 1994）在研究6~9岁儿童的亲社会行为时，检验了与儿童助人行为倾向（如，捐款给有需要的儿童）相关的心理因素。他们发现一些变量——如同情心、情感推理和关于金钱的知识等——单

独作用时，它们和亲社会行为之间的相关很低。但是，当这些变量联合作用时，能够很好地预测亲社会行为。例如，具有较强的同情心、较强的情感推理并对金钱有所认识的儿童，捐款的数目是在这些变量上表现较低的儿童的4倍。

发展心理学家丹·基廷（Dan Keating, 2007）评论了一篇有关美国一些州的毕业生驾驶执照项目和青少年驾车安全之间关系的文献。这些项目确实起到了作用——它们降低了青少年驾驶撞车和死亡的概率。然而，在不同的州结果是不一样的，每个州在几个基本部分之外都有各自不同的细则，如需要驾驶培训、乘客限制、夜间驾驶限制、法定年龄、最小练习时长、初学者许可时间。因此，问题就变成这些组成中每个成分是否真的有效，以及他们是否存在交互作用。研究表明，没有一个成分能降低青少年撞车或死亡的概率。但是，他们组合在一起能使青少年死亡率降低超过20%。

因此，原因多样化的概念可能比你最初设想的要复杂得多。不仅需要追踪并测量影响问题行为的种种可能因素，还必须考察这些变量是如何共同作用的。

临床心理学家斯科特·利连菲尔德（Scott Lilienfeld, 2006）对变量的因果影响进行了讨论，认为其是一个从强到弱的连续体。只有当一个变量处于这一连续体的最强端时，它才能独立产生作用。因果影响的最强形式，是一个变量对于其对应变量产生的影响来说既是必要的又是充分的。变量必须出现效应才会产生（必要性），对变量来说，其自身必须充分到能够产生效应。较弱形式的影响则涉及一个变量和其他变量的关系。一个原因变量可能是必要的（因变量表现出效应时原因变量必须存在）但并非充分的（它依赖于其他变量出现才能产生效应）。最后，一个弱的原因变量可能既不是充分的也不是必要的——它的出现只是增加了效应的整体上的统计概率。

单一原因解释的诱惑

复杂事件是由多重原因所决定的，这个基本的理念似乎很容易理解。实际上，当问题没有太大争议时，这个观点确实很容易掌握和运用；但是，当预设偏见——这个科学工作者的老敌人（参见第3章）开始抬头时，人们就会倾向于忘记原因多样性这一原则。我们无数次听到，人们在争论犯罪原因、财富分配、恐怖主义的成因、妇女和少数族裔待遇、贫困、死刑的作用以及纳税标准这类容易引发情绪的话题时，其方式都是让人觉得这些问题是简单的、单维的，而且导致结果的原因只有一个。如果直接询问多种原因，人们有时会承认原因多样性的存在；但是他们很少会自发地对他们关心的事情提供出不同的原因作为解释。很多时候，人们习惯用“零和”态度去对待潜在的原因——所有的原因都和其他的原因竞争，强调一个的必要性会降低另一个的重要性。“零和”观点是错误的。

“零和”游戏——一个人的收益是另一个人的损失——常常反映了我们如何讨论那些容易引发情绪的话题。在情绪的影响下，人们通常会忘记原因多样性这一原则。想想两个敌对的政党是如何讨论社会犯罪问题的。自由主义者会认为那些社会经济地位低下的人之所以会犯罪，是因为他们本身就是恶劣社会环境（如失业、恶劣的住房条件、缺乏教育和对未来丧失希望等）的受害者。而比较保守的人会争辩说，也有许多穷人并没有犯罪，所以社会经济条件并不是主要原因。与之相反，他们认为个人的价值观和人格特征才是决定犯罪行为的真正原因。双方似乎都没有认识到个体因素和环境因素共同导致了犯罪行为。犯罪没有单一原因的解释。犯罪行为由多种因素决定，其中一些是环境方面的，而另一些是个体属性方面的。

再看一下有关复杂经济结果成因的讨论。这些结果都是由多种因素决定的，因而难以精确地进行预测。例如，经济学的争论聚焦在一个过去几十年来具有重要社会意义的问题上：美国贫富差距的扩大（Bartels, 2008; Bilmes & Stiglitz, 2009; Brooks, 2008; Gelman, 2008; Madrick, 2006; Surowiecki, 2010）。在这里，事实并不存在争议，对事实的解释才是争议的主题。事实是这样的：自从1979年以来，所有美国男性劳动者的真实收入（如根据通胀率调整过后的）一直停滞。因此，美国中等收入家庭和低收入的家庭都几乎无法用收入

养活自己。与之相反，占美国人口数1%的顶尖高收入者，同期收入（根据通胀率调整过后的真实收入）增长超过100%。另一个解释是，美国1980到2005年间超过80%的收入都到了最富的1%的纳税人手里（Bartels, 2008）。在1977年，占人口20%的最富者的收入是20%最穷者的4倍，到2006年时已经超过10倍。

财富从公民的一个阶层大规模地转移到另一个阶层手中，这一现象引发了一场极富争议的、有关其原因及影响的政治辩论。这场争辩最引人注目之处就是，这些争论者都只关注单一的原因。争辩的每一方都只以某一个原因为立论基础，然后千方百计地攻击所有支持其他原因的观点。事实上，计量经济学研究（Bartels, 2008; Bilmes & Stiglitz, 2009; Gelman, 2008; Madrick, 2006）已经聚焦了四个变量（还有人提出了超过四个的变量，但这四个是得到最广泛的关注及研究的）：第一个因素是新移民不断涌入美国，而这些人多是非熟练工，他们造成了非熟练劳动力供大于求，使得已经很低的工资水平继续下滑；第二个原因是全球化，它进一步加剧了收入不均，因为公司可以通过业务外包，在一些工资水平较低的国家雇用一些非熟练工和半熟练工（正变为熟练工），而这更加重了本国非熟练劳动力的过剩；第三个原因是工会和大企业在力量对比方面的此消彼长；第四个因素是1980年和2001年两次减税错误地减轻了富人的税赋。

经济学研究这四个变量的时候到底发现了什么呢？你已经猜到了。所有这四个因素共同作用造成了不断加重的社会不平等。这个例子也证明了先前所提到的交互作用的概念。重要的是，所有有关于此的研究都指出，这些因素存在交互作用并彼此强化。更为激烈的全球化竞争使得企业在与工会的斗争中拥有更大话语权，与此同时，移民降低了非熟练劳动力的比例，也在一定程度上让现有的工会更加难以讨价还价。

和经济学的问题一样，心理学所研究的几乎所有复杂问题也都是由多重原因决定的。以学习障碍为例，这个问题已经被教育心理学家、认知心理学家和发展心理学家广泛地研究过。结果发现，脑部的病变与学习障碍有关（Shaywitz & Shaywitz, 2004; Snowling & Hulme, 2005; Tanaka et al., 2011; Wolf, 2007）。还有研究发现，学习障碍具有遗传方面的原因（Olson, 2004; Pennington & Olson, 2005）。这两个研究结果看起来好像可以让我们做出一个结论：学习障碍是纯粹的生理—脑的问题。但这样的结论是错误的，因为进一步的研究发现，造成学习障碍的部分原因是在早期学校教育中缺乏某些指导性的经验（Pressley, 2002），以及贫困的家庭环境（Dickinson & Neuman, 2005; Senechal, 2006）。学习障碍因此不是由单一原因

所引起的；相反，它是生理与环境因素交互作用的结果。

在抑郁的产生原因和治疗方法方面也存在类似的情况。抑郁是由基因偏转性和环境风险因素共同决定的。与之相似，在对抑郁的治疗上，“医学加心理疗法”看似能够产生最佳的治疗效果（Engel, 2008）。

一旦找到了复杂现象的多重原因，且这个现象是一个亟待解决的问题，这就意味着问题的解决需要多重的干预。数十年前，我们有一个主要的健康问题——吸烟行为非常普遍，而吸烟与多种疾病有关。近几十年，多种不同的干预方式都降低了全社会的吸烟水平，如禁止播放烟草广告、提高烟草税、让尼古丁贴片随处可得、公众场合禁止吸烟等（Brody, 2011）。慢慢地，由于针对多种原因的多重干预起到了效果，吸烟率下降了。

就像有许多干预措施减少吸烟一样，需要采用多重社会干预才能使当前全国性的肥胖问题得到控制和扭转（Chernev, 2011; Herman & Polivy, 2005）。原因是我们目前肥胖的流行在几十年之前就开始了，诸多不同的社会潮流共同导致了这种结果：居住在郊区减少了散步；更多女性进入职场使得在家做饭的机会减少；快餐产业爆炸式增长；食品广告无所不在；电子游戏使儿童久坐不动以及其他很多因素（Brody, 2008, 2011）。相应地，解决这个国民性问题的方法也应当是多方面的。

最后举一个例子，假设招募了一些心理学家让他们运用知识去推断恐怖主义的成因（Kruglanski, Crenshaw, Post, & Victoroff, 2007）。由于原因是多重的，解决的方法也是多维的。例如，对于帮助被捕的恐怖分子消除过激行为，心理学家推荐出许多干预方式，包括：与家人一起工作；鼓励婚姻的计划；职业训练；启用在宗教对话中专家身份的学者，等等（Kruglanski, Gelfand, & Gunaratna, 2010）。

小结

本章内容虽然简单，但却非常重要。考察行为的原因时，要依照多样性的原则来思考。不要陷入误区，认为某一特定行为只是由某一特殊原因造成。大部分复杂的行为都是由多重原因所决定的。各种各样的因素共同起作用才导致了某种行为出现。有时多个因素联合在一起时会产生交互作用。也就是说，变量共同作用时的整体效应，会和其单独作用时获得的效应完全不同。

Chapter 10
人类认知的阿喀琉斯之踵：概率推理

问：男人比女人高，对吗？

答：“对。”

问：所有男人都比所有女人高，对吗？

答：“错。”

完全正确。信不信由你，在这一章里，我们还将花一些篇幅来讨论一些问题，从你刚才对上面两个问题r回答可以看出，你已经知晓了一些答案。但是，先别因此就跳过这一章。因为接下来，在我们对一些看似非常简单的原则所作的解释之中，会有惊喜等着你。

你为第一个问题给出了肯定的答案，这是因为你没有把“男人比女人高”这句话理解成第二个句子所说的“所有的男人都比所有的女人高”。你把第一句问话正确地理解为“男人有比女人高的趋势”的意思，因为每一个人都知道，不是所有的男人都比所有的女人高。你理解到那句问话反映了一个概率趋势，而不是一个在任何情境中都适用的事实。我们所说的概率趋势是指有较大的可能性，但并非在所有情况下都必然如此。也就是说，性别和身高的关系要用可能性和概率的词汇来描述，而不是用必然性的字眼。在自然界中，很多关系的本质也是概率性的，例如，接近赤道的地区比较热；每家的孩子数目不超过8个；地球上大部分地区昆虫的数量比人类多。这些都是统计学可证明的趋势，但是它们当中的每一句话都不是绝对的，仍然可能会有例

外。因为它们是概率的趋势和规律，而不是在所有情况下都成立的关系。

2008年夏天，广受欢迎的58岁政治播音员蒂姆·拉瑟特（Tim Russert）死于心脏病，这给美国人在医学的概率性知识方面上了悲伤的一课。拉瑟特服用胆固醇药片和阿司匹林，骑运动自行车，每年都进行压力测试，但是他仍死于心脏病（Grady, 2008）。他对健康极为关注，这让很多《纽约时报》的读者写信说一定是医生遗漏了什么。这些读者不明白医学知识也是概率性的。每个失败的预测并不是错误。事实上，他的医生没有遗漏任何事。他们最大限度地应用好自己的概率性知识，但这不意味他们能预测每例心脏病。科学作家丹尼斯·格雷迪（Denise Grady, 2008）告诉我们，根据拉瑟特先生最后一次压力测试和其他状态的诊断，医生根据一个广泛使用的公式估计拉瑟特在十年内罹患心脏病的概率为5%。这意味着和拉瑟特先生相似的人中，100人中有95人十年内都不会得心脏病。拉瑟特先生恰恰是那不幸的5个人中的一个。概率性的医学科学不能提前告诉我们谁是那不幸的5个人。

蒂姆·拉瑟特的例子提供了一个契机，能够让我们强调概率预测事实上就是真正的预测。这就是我们想说的。因为概率预测是数值性的，因此是抽象的，人们有时很难把它当作一个真实的事物。因为不能预先知道那“百分之五”姓氏名谁，人们总是觉得预测不会像它本来的那样真实。但是在他们死了之后，这个“五”就有了具体的名字。蒂姆·拉瑟特就是五人之一。即使我们提前就将他命名为死者，他死亡的概率也不会因此减少。我们必须克服由于数值抽象性引发的“概率预测不是真实的”这种感受。

科学家们在进行概率预测时，实际上是在讨论那些活生生的人。回顾一下第8章的观点，由于在开车的时候打电话和发短信，数以百计的美国人在撞车事故中无谓地死去。由于这是概率预测，所以我无法说出到底这数以百计的人都是谁。然而，这个预测的真实性并不因为它是概率性的而降低。或许我们可以用一种更生动的方式表明这一点：在读了“由于司机开车时打电话分心导致了一场车祸”的报道之后，某些人今年的命运改变了。

人们很难接受概率性预测的现实——人们并不是生活在一个确定的世界中。科学作家纳塔利·安吉尔（Natalie Angier, 2007）讨论了一个问题：人们认为地质学家能够预测每一次地震，但是为了不引起恐慌因而他们不对外公布消息。一个地质学家曾收到一个女人的来信，请他将自己儿子送到城外亲戚家时告诉她一声。凭借这个例子，安吉尔指出，人们似乎更倾向于认为权威都致力于巨大的谎言，而不

是简单承认科学的不确定性。吉仁泽和他的同事之前的研究（Gigerenzer et al., 2007）确认了安吉尔的担忧。吉仁泽发现在一个德国城市的人群样本中，44%的人（错误地）认为乳腺X光检测给出的是一个“绝对确定”的结果，63%的人（错误地）认为指纹识别也是一个“绝对确定”的结果。

事实上，心理科学所揭示的所有事实和关系都是用概率来表述的。这一点也并非心理学所独有。在其他学科里，很多定律和关系也是用概率而非必然性来表述的。例如，人口遗传学的所有子学科都基于概率关系。物理学家告诉我们，原子中电子负荷的分布也是通过概率函数来描述的。因此，各种行为关系都是以概率形式加以描述的，然而这一事实并没有使得它与其他科学之间产生天壤之别。

很多作家都指出：“人们似乎生活在有时和或许的世界里，但他们希望继续生活在永远的确定中。”（Bronowski, 1978a, p.94）在这一章里，我们想尽可能地让你在这个“有时和或许的世界”里感到更舒服一些，因为一个人若想要理解心理学，就必须对“概率推理”这一本章的主题安之若素。

“某某人”统计学

大部分公众都能意识到，医学的许多结论都采用的是概率趋势而非绝对确定性的表述。吸烟会导致肺癌并诱发其他健康问题，相关的医学证据汗牛充栋（Gigerenzer et al., 2007）。但每个吸烟者都会得肺癌吗？所有戒烟者都解除了患肺癌的风险吗？大多数人都不会认为这些推论能够成立。吸烟很大程度上增加了患肺癌的概率，但并非绝对。医学能够以很大的把握告诉我们，吸烟群体中的人比与之相似的非吸烟群体中的人更容易死于肺癌，但不能告诉我们是哪一些人会死，这种关系就是概率，它并不适用于所有个案。我们都知道这一点——真的知道吗？我们经常看到下面这样的场景：一个不吸烟的人引用吸烟导致肺癌的统计数据，试图说服一个瘾君子戒烟，所得到的结果仅仅是对方的反唇相讥：“嘿，走远点儿！你看那个铺子里的老乔，他从16岁开始，每天要吸三包骆驼烟！现在他已经81岁了，看上去还很硬朗！”人们对此可能作出的推断显而易见——就是这一个特例已经推翻了吸烟和肺癌之间的关系。

令人吃惊和沮丧的是，这种反驳手段屡试不爽。通常情况是，每当个别个案被用来证明概率趋势是无效的时候，很多人都常常点头表示赞同，这反映出他们没有正确理解统计规律的本质。如果人们认为一个特例就可以让一个规律失效，他们一定认为这个规律应该在任何情况下都适用。

简而言之，他们错误理解了概率定律的性质。即使是最强的趋势也会有少数的“特例”与之相悖。就拿吸烟的例子来说，活到85岁的人中只有5%是吸烟者（University of California, Berkeley, 1991）。或者从另一角度来看，活到85岁的人中有95%属于从不吸烟者，或在一段时期内吸烟但最终戒断者。连续从未间断地吸烟会显著地缩短寿命，然而也有少数吸烟者活到了85岁。

心理学家把类似“老乔”的故事称作“某某人”统计学的运用：由于某些人知道与某个成熟的统计学趋势相左的“某某人”的例子，就会质疑这个趋势。例如，我们经常听到类似的话——“你是说服务业的就业机会正在扩大而重工业中则在缩小？这不对，我就知道‘某某人’上周四在一个钢铁厂找到了一份工作”；“你说与30年前相比，家里的孩

子少了？少胡扯！隔壁的年轻夫妇已经有了3个小孩，但他们还不到30岁”；“你说通常孩子都会倾向于信仰他们父母所信仰的宗教？但据我所知，我的一个同事的孩子就在前几天改信了另一门宗教。”

当我们面对和过去持有的观念相矛盾同时又是强有力的证据时，无所不在的“某某人”总是会立刻跳出来否定这些统计规律。因此，我们可以说，实际上人们知道的不少，他们只不过顺手把“某某人”当成一种工具，把与他们观念相悖的事实否决掉而已。然而，研究人类决策和推理的心理学家们的研究结果表明，人们之所以使用“某某人”，不只是由于它是一个有用的辩论手段。相反，这一错误的争论模式之所以被应用得如此频繁，主要在于人们不知道如何处理概率信息。决策心理学的最新研究发现，概率推理可能正是人类认知的阿喀琉斯之踵。

概率推理以及对心理学的误解

由于人们在运用概率信息方面存在问题，导致心理学的研究结果常常被误解。我们都理解“男人比女人高”是一个概率趋势的陈述，并不会因为有一个特例（某个男人比某个女人矮）就认为这一陈述是错的。很多人也能以同样的方式来理解“吸烟可以导致肺癌”的陈述（尽管对于那些不愿相信吸烟会导致其丧命的瘾君子们来说，“老乔”可能还是有说服力的），然而，与之相似的有关行为趋势的概率表述却引发了广泛的猜忌，而且常常是“某某人”刚一露头，这种概率表述便被人们抛弃了。很多心理学教师在讨论某些行为之间关系的证据时，都往往得到同样的反应。例如，教师可以呈现如下的事实：儿童的学业成绩和家庭的社会经济地位及父母的教育水平相关。但这个事实常常会遭到至少一个学生的反对，他会说，他有一个朋友是国家优秀奖学金获得者，但是他的父亲只是中学毕业。甚至那些理解吸烟—肺癌例子的人，对这一问题的态度也变得摇摆不定了。

人们从没想到过要用“某某人”的论据来反驳医学和物理上的发现，却习惯于用之驳斥心理学的研究结果。大多数人能理解医学科学提出的治疗、理论及事实是概率性的。例如，他们理解一种药对一组病人来说，并不是对他们个个都有疗效，而且医学也经常不能事先告诉我们，该药会对哪些病人有疗效。通常可以说，100个病人接受某治疗方案，100个病人不接受任何治疗，在一段时间之后，接受治疗的这100个病人总体来说会比不接受治疗的100个病人的病情好转一些。在前面的章节中，我提到过我一直服用一种叫做舒马曲坦（舒马曲坦琥珀酸盐）的药物以缓解偏头痛。说明书上的信息告诉我：“控制研究证明，在一个特定的剂量水平上，57%的服用此药的患者在2小时内得到症状缓解。”我就是那幸运的57%中的一个，但是制药公司和我的医生都不能保证我不会是那不幸的43%。药物并不是对每个患者都起效的。没有人因为这个并非在所有情况下都适用的概率表述，就怀疑这一治疗的价值。许多心理学的研究结果及心理治疗的效果也存在类似的情况。然而，一旦心理学研究结果和心理治疗效果不能在所有情况下都适用，就常常会引起人们对心理学产生极大的失望和蔑视。一旦面对心理学的话题，人们常常忘记一个最基本的原则，那就是知识不需要完全确定后才是有用的——即便某些知识不能预测

个体的具体情况，但如果能对群体的总体趋势有预测能力，也是非常有益的。基于群体的特征所作的结果预测常常被称为总体统计数字或统计预测（下一章将详细讨论统计预测这一概念）。

想想看，当一个不健康的人去看病，医生说除非他进行锻炼和改变饮食习惯，否则会有很高的风险发作心脏病。我们不会因为医生没有告诉这个人“如果不改变饮食习惯，他将于2014年9月18日心脏病发作”，而认为医生的信息是无用的。我们容易理解该医生的预测是概率性的，并不能达到那种精度。同样，当地质学家告诉我们，某地区在未来30年发生一场震级为8.0或更大地震的可能性为80%时，我们不会因为他们没有说“2016年7月5日就会有地震发生在这里”而贬低其知识。科学作家伊丽莎白·柯尔伯特（Elizabeth Kolbert, 2005）讲述了一群杰出的气象学家如何为了教育公众而在他们的网站上贴出一篇文名为“新奥尔良会不会是第一个由于人类造成的气候变化而被毁灭的美国城市”的文章，他们指出标题的问题完全是错误的。他们说“全球变暖和单独一次的飓风（或干旱、炎热、洪水）没有任何关系，只和大的数据模式有关”（p. 36）。

当学校心理学家推荐一个针对学习障碍儿童的训练计划时，显然是在作概率预测——该训练能使这些儿童有较大的可能性获得好成绩。当一个临床心理学家推荐一个针对有自我伤害行为的孩子的计划时，情况也与之类似。心理学家判断如果按计划进行治疗，会有较高的概率获得一个很好的结果。但是不同于心脏病发作和地震的例子，心理学家常常要面对诸如“但我的孩子何时能达到某一年级的阅读水平”或“他在这个治疗计划中要待多久”这类问题。这些问题都是无法回答的，正如地震和心脏病何时发生也是无法回答的一样，因为针对所有这些问题——心脏病发作、学习障碍儿童、地震以及自我伤害的儿童——所作的预测都是概率性的。

出于这些原因，全面认识概率推理对理解心理学至关重要。耐人寻味而又颇具讽刺意味的是，心理学很可能是人们不能进行统计思维的最大受害者，然而在所有学科中，心理学是对人类概率推理能力研究最多的学科。

有关概率推理的心理学研究

过去的30年里，普林斯顿大学的丹尼尔·卡尼曼（Daniel Kahneman, 2002年诺贝尔奖得主）及已故的阿莫斯·特维斯基（Amos Tversky）等心理学家的研究，彻底改变了我们对人类推理能力的认识。他们在研究中发现，很多人头脑里压根儿没有概率推理的基本原则，更多的人则是有一些但并不完备。正如学者经常指出的，这些基本原则在人们头脑里没有充分发展并不足为奇。作为数学的一个分支，概率论是最近才发展起来的（Hacking, 1975）。概率论中最关键的发现直到16和17世纪（Mazur, 2010）才产生，并且许多重要进展的产生也就是20世纪的事。

概率论的关键进展产生的年代意味着一个重要的事实：在概率定理被发现之前，机遇游戏已经存在了好几个世纪了。这又是一个例证：个人经验不足以让人们获得对世界的基本理解（参见第7章）。针对概率定律的正式研究发现了机遇游戏的运作机制，而成千上万的赌徒以及他们的个人经验，并不足以揭示机遇游戏的本质。

问题在于，社会越复杂，人们就越需要概率思维。如果一个普通人想要对其生活的社会有一个基本的理解，那么，他至少应具备统计思维这一最基本的能力。

你或许有以下疑问："为什么他们要提高我的保险费？为什么张三的保费比李四高，是不是社保局穷疯了？我们州的彩票有黑幕吗？犯罪率到底是在增加还是在减少？为什么医生要安排这些检查？为什么欧洲人可以用一些很珍稀的药，而美国人就不行？做相同的工作，女性赚的真的比男性少吗？国际贸易真的减少了美国人的就业机会，并降低了他们的薪酬吗？日本的教育要比我们好吗？加拿大的卫生保健真的比美国好且价格低廉吗？"这些问题都问得很好，这都是关于我们的社会如何运作的具体而实际的问题。要知道每个问题的答案，我们就必须运用统计思维。

显然，本书由于篇幅所限，不可能全面讨论统计思维。然而，我们将简要地讨论某些概率推理中的普遍误区。学习概率思维技巧的最好方法就是察觉人们在统计推理时最常犯的错误是什么。

对概率信息的不充分利用

在心理学领域中，有一个已经被反复证实的发现，那就是一个具体事件的信息往往可以完全击败较为抽象的概率信息（第4章中讨论的“鲜活性”问题）。忽视概率信息的例子比比皆是，而且并不仅仅局限于缺乏科学知识的外行人。这有一个连有经验的医生都容易做错的问题：如果在每1000人中有1人携带艾滋病病毒（HIV），再假设有一种检查可以百分百地诊断出真正携带该病毒的人，最后，假设这个检查有5%的阳性误诊率。也就是说，这项检查在没有携带HIV的人中，也会错误地检测出有5%的人是病毒携带者。假设我们随便找一个人来进行这项检查，结果呈阳性反应，表明此人为HIV携带者。假定我们不知道这个人的患病史，那么他真的是HIV携带者的概率是多少呢？

普遍的回答是95%，正确的答案是约2%。医生们过分高估了阳性结果表示患病的概率，因为他们一方面过分重视个案信息，另一方面又忽视了基础比率信息，从而过高地估计了阳性检测结果所真正代表的患病概率。稍稍进行逻辑推理就可以说明基础比率对概率的重要作用。1000个人当中只有1人是真正的HIV阳性者。如果另外999人（不患病）也进行了此项检查，由于这一检查有5%的虚报率，他们当中将有接近50人（999乘以0.05）会被检查出携带这种病毒。这样一来，呈阳性反应的人就会是51个。因为在这51个人当中，只有1人是真正的HIV阳性者，此人确诊得病的概率其实只接近2%。简而言之，基础比率就是绝大多数人没有携带这种病毒（病毒携带者只有千分之一）。这个事实和确定的虚报率综合考虑，就能使人确信，在绝对数量上，大部分呈阳性反应的人并不携带这种病毒。

尽管大多数人很快就意识到了以上概率逻辑的正确性，但他们最初的直觉反应却是忽视基础比率，并过分看重临床检测的证据。简单来说，事实上人们知道什么是对的，但却本能地做出了错误结论。心理学家把这类问题称为认知错觉（参见Kahneman, 2011; Pohl, 2004）。在认知错觉中，即使人们知道正确答案，他们也会由于问题的问法不同而做出错误的结论。

在这一问题里，个案证据（实验室的研究结果）好像是摸得着的、具体的，而概率证据则好像是摸不着、不确定的。当然，这种理解是错误的，因为个案证据本身一定是概率性的。一项临床检验会以一定的概率对疾病作出误诊。上述情境就是一个例子，要想做出正确的决策，就必须结合考虑两种概率——对个案证据作出正确或错误诊断的概率（即95%或5%）和过去经验所提供的先验概率（也叫基础

比率）。整合这些概率的方法有的是正确的，也有的是错误的，并且时常是错的——特别是当个案证据给人一种很具体的错觉时（请回忆在第4章所讨论的鲜活性问题），人们往往会以错误的方式来整合信息。

上面HIV的例子也说明，当解释测试结果的时候要注意假阳性率。在例子中，较大的假阳性率（5%）和HIV很低的基础概率（只有千分之一）相结合导致了以下结果：检验出阳性的人没患病的可能性要比患病的可能性高。假阳性问题在诊断测验中非常受重视，尽管在医学上治疗和诊断上有巨大的进步，但是临床测验还是存在高假阳性率。在一个针对3000名老人的研究中，发现在进行前列腺、肺、结直肠癌的筛选测试后，超过三分之一的男性都出现假阳性的结果。这项测验表明，当他们没有癌症的时候，被测出了癌症（Croswell, et al., 2009）。

样本大小信息的误用

请大家思考下面两个由特维斯基和卡尼曼（Tversky & Kahneman, 1974）提出的问题。

1. 一个小镇里有大小两所医院。在大医院里，每天大约有45个婴儿出生；在小医院里，每天大约有15个婴儿出生。如你所知，大约有50%的婴儿是男孩，但具体的百分比每天都不一样，有时候高于50%，有时候低于50%。每一所医院都记录了一年内出生的男婴比例高于60%的天数。你认为哪一所医院记录的天数多？

a. 大医院；

b. 小医院；

c. 基本一样。

2. 假设一个容器里装满了球，其中有2/3是一种颜色，其余1/3是另一种颜色。一个人从中拿出5个球，发现有4个是红色的，1个是白色的。另一个人从里面拿出20个球，发现有12个是红色的，8个是白色的。哪一个人会更自信地认为这个容器里有2/3的球是红色的、1/3的球是白色的，而不是有1/3的球是红色的、2/3的球是白色的？这两个人会给出什么样的概率呢？

对于第一个问题，大多数人回答“基本一样”。剩下的人则一半选

择大医院，一半选择小医院。但正确的答案是小医院，所以接近75%的被试都给出了错误答案。答错是由于人们没有认识到样本大小在这个问题中的重要性。当其他因素保持不变时，较大的样本总是能够更精确地估计出总体的真正数值。也就是说，在任何一个指定的日子，大医院由于有较大的样本，男婴出生的概率更趋近于50%。相反，小的样本总是倾向于距离总体平均值比较远。因此，小医院将会有更多的天数记录了与总体平均值相矛盾的男婴比率（60%，40%，80%，等等）。

在回答第二个问题时，大多数人认为5个球的样本提供了更令人信服的证据，能证明这个容器里的球大多数是红色的。事实上，概率恰恰与之相反。对5球样本来说，坛里大部分为红球的概率是8:1。而在20个球的样本中，这个几率是16:1。尽管在5个球的样本中，抓出红球的比例较高（80%:60%），但考虑一下，另一个样本的大小是其4倍，因此对球的比例能够做出更为精确的估计。然而，大部分被试被5球样本中红球有较高的比例给迷惑了，而没有充分考虑到20个球的样本具有更大的可信度。

这两个例子证明了有关样本量的一个非常有用的原则：样本越小，产生极端值的可能性就越大。心理学家丹尼尔·卡尼曼（2011）向我们举了一个例子，在因果研究中如果不能应用这一原则，只能使我们白费劲。他指出，美国一项关于3141个地区的研究发现，肾肿瘤发病率最低的基本上是人口稀疏的乡村地区。卡尼曼（2011）指出，这很容易产生一个因果理论去解释为什么会这样：“乡村干净的生活环境没有空气污染、没有水污染、食物没有食品添加剂。”（p.109）这个因果理论唯一的问题是，它没有解释在相同研究中的另一个发现：肾肿瘤发病率最高的地区也通常是人口稀疏的乡村地区！如果我们最初知道的是后一结果，我们开始可能会作出这样的解释：乡村的人抽更多的烟、喝更多的酒、有更高脂肪的饮食习惯。但是这个解释以及低发生率的解释都不准确。这就是我们之前讨论的医院例子的现实版，人口稀少的乡村地区是一个小样本，因此必然会产生各类极端值——极端高值和极端低值。

许多人无法理解到他们所处的情境都涉及样本量。也就是说，他们很难认识到他们处理的是样本，而不是全部实体。无法意识到这一点会导致忽略这样一个事实：样本测量受到取样误差的影响。例如，你的医师给你验血，从你身上取走的血是一个样本，并且对它而不是你整个血液系统进行检验。假设这个样本代表了你的整个血液系统，但是假设是概率性的，并且只是或多或少是真实的。这其中会产生一些误差，由于测试不能测试全部的血液系统，因此样本中的细胞以及

成分性质必定会和绝对真实有所偏离。简言之，你的医师是基于你很小的血液样本而作出了有关你血液全部成分的推断。

肿瘤的活体组织切片检查也与之类似。由于活组织切片只是更大肿瘤上的一个小样本，因此也存在误差。医学作家塔拉·克-波普（Tara Parker-Pope, 2011）在讨论疑似前列腺癌的活体组织切片时告诉我们，常规的活组织切片样本只有前列腺的千分之一点三。她提出证据表明在20%的样本中会发生阶段和等级方面的错误。关键是要认识到，在进行行为测量时也是一样的。我们经常用一个小样本代表一个大得多的群体的行为。

在不同领域中进行证据评估时需要遵守的一条基本原则，就是认识到样本规模对信息可信度的影响，这对于理解行为科学的研究结果尤为重要。不管我们是否意识到，我们会对较大的群体持有一些普遍的看法。我们很少察觉到，我们最坚定的信念是建立在多么脆弱的事实基础之上。把对几个邻居和同事的观察以及在电视新闻上看到的一些趣闻轶事放在一起，我们就迫不及待地要对“人性”或者“美国人”发表见解。

赌徒谬误

请回答下面两个问题：

问题A：想象一下你在掷一枚普通的硬币（硬币出现正面和反面的概率各占50%），已经连续出现了5次正面。对于第6次，你认为

_____出现反面的概率比正面要大

_____出现正面的概率比反面要大

_____正面和反面出现的概率一样大

问题B：玩老虎机的时候，赢钱的机会是1/10。茱丽头3次都赢了。她下次赢的几率是____分之____

这两个问题是为了检测你是否容易出现所谓的赌徒谬误——即倾向于将过去事件和未来事件之间联系起来，而实际上两者是独立的。两个结果是相互独立的，一个事件的出现不会影响另一事件出现的概率，大多数机遇游戏都具备这种性质。例如，幸运轮盘的数字与之前的数字无关。轮盘数字一半是红的，另一半是黑色的（为简化起见，

我们将忽略绿色的零和双零），所以对任意一次旋转来说，出现红色的概率均等（0.50）。然而在连续5~6次出现红色数字之后，许多投注者转投黑色，因为他们认为现在黑色更有可能出现。这就是赌徒谬误：明明是独立事件，却认为先前的结果会影响下一结果出现的概率。在这种情况下，投注者错在他们的信念。轮盘并不记得先前发生过什么。即使连续出现15个红色数字，红色数字在下轮出现的概率仍然是0.50。

在问题A中，有些人认为在5次出现正面之后，反面更可能出现。他们这么想就陷入了赌徒谬误。正确的答案是，正面和反面在第6次中出现的可能性一样大。同样，对问题B任何非1/10的回答都落入了赌徒谬误。

赌徒谬误不仅限于没有经验的赌徒。研究表明，即使是那些一周赌20小时的资深赌徒，仍然表现出赌徒谬误（Petry, 2005; Wagenaar, 1988）。事实上，研究表明，正在接受赌博脱瘾治疗的个体比对照组更相信赌徒谬误（Toplak, Liu, Macpherson, Toneatto, & Stanovich, 2007）。

重要的是我们要认识到，这一谬误不仅限于赌博游戏，它还存在于任何概率起着重要作用的地方。换句话说，它几乎无处不在。婴儿的基因构成就是一个例子。心理学家、医生和婚姻顾问常常遇到一些已有两个女孩的夫妇，他们正计划要生第三个孩子，因为“我们想要个男孩，这回一定是个男孩”。这就是赌徒谬误，在生了两个女孩之后生男孩的概率（接近50%）和生第一个孩子时完全一样。生了两个女孩不会增加第三个孩子是男孩的概率。

赌徒谬误来源于对概率的诸多错误认识。其中一个错误认识就是，如果一个过程真正是随机的，就不可能出现重复同一结果或某种模式的序列，哪怕是一个不起眼的随机事件（例如，掷6次硬币）。人们习惯性地低估了重复（正正正正）或某种模式（正正反反正正反反正正反反）在一个随机序列中出现的可能性。正因为如此，人们在模拟一组真正的随机序列时，常常适得其反地产生出一个很少出现重复和某种模式的排列。这是因为人们往往会错误地让可能的结果尽量轮流出现，以为这样才称得上是随机抽样，这无疑破坏了真正的随机排列中可能出现的结构（Olivola & Oppenheimer, 2008; Scholl & Greifeneder, 2011）。

那些声称自己有通灵能力的人可以轻而易举地利用人们的这一错觉。大学心理学课上常会进行这样一种演示，老师让一名学生准备200个数字的排列，这200个数字从1、2、3这三个数字中随机重复抽

取。完成之后，不要让老师看到。接下来，让这名学生全神贯注于他写的第一个数字上，老师则来猜这个数字是什么。当老师说出他的猜测之后，这个学生再向全班同学及老师公布正确的答案。有人记录猜对的次数，直至猜完这200个数字。在实验开始之前，这个老师声称有通灵能力，可以在实验过程中用读心术来证明“通灵能力”的存在。通常在展示之前，老师会先问班里的学生，他猜测的成绩要达到多少——也就是“猜中”的百分比是多少才算是能证明他确实有通灵能力。这时，通常都会有一个修过统计课程的学生回答说，因为纯粹随机的猜测也能猜中33%，所以要想让别人相信他有通灵术，猜中的比例就一定要超过33%，至少达到40%。班上大部分同学都会认同这一个观点。演示结束后，结果那位老师猜中的比例果真超过了40%。这个结果令很多同学感到惊讶。

学生们从这一演示中领教了什么是随机性，并且知道伪装通灵能力是多么地容易。在这个例子中，老师仅仅利用了“人们不让连续重复的数字出现”这一事实：人们频繁地在三个数字间换来换去以制造“随机性”。在真正的随机序列中，已经出现了三个2之后，再出现2的概率是多少呢？其实还是1/3, 与出现1或3的概率一样大。但大多数人在产生随机数字时并非如此。出现一个哪怕很小的重复片断之后，人们也常常会刻意地变换数字，力图制造一个“随机”序列。这样，在我们的这个例子中，老师只要在每一轮猜测前，不去挑选那个学生在前一轮中挑选的那个数字，而从另外两个数字中选一个就可以了。例如，如果那个实验中的学生在上一轮说的数字是2, 那么老师就会在下一轮的猜测中从1或3中任选一个。如果学生在上一轮说的数字是3, 那么老师就会在下一轮的猜测中从1或2中任选一个。这样一个简单的把戏根本不需要什么通灵能力，就能保证猜中的概率高于33%——高于三个数字随机猜测的准确率。

人们总是认为，如果一个序列是随机的，那它就不应呈现有重复和某种模式。2005年关于美国苹果公司出品的数码音乐播放器iPod“shuffle”模式（意即“随机播放”）的争议就以一种幽默的方式证明了这一点（Levy, 2005）。此模式将下载到iPod里的歌曲以随机方式播放。很多用户抱怨说shuffle模式并不随机，因为他们经常听到同一专辑或同一曲风的歌曲。当然，许多心理学家和统计学家在听到这类抱怨时只能暗自苦笑，因为他们了解我刚才提到的类似研究。科普作家史蒂芬·列维（Steven Levy, 2005）讲述了他经历过的类似事情。他的播放器似乎在起初的一个小时里偏爱史提利·丹（Steely Dan）的歌！但列维明智地接受了专家告诉他的事实：真正的随机序列，往往看起来不像是随机的，因为人们倾向于在所有的地方看到固定模式。

再谈统计与概率

以上列举的涉及统计推理理解中出现的错误，仅为冰山一角，有可能阻碍人们正确理解心理学。有兴趣的读者可以阅读由吉洛维奇（Gilovich）、格里芬（Griffin）和卡尼曼（Kahneman）编写的《启发式和偏见：直觉判断的心理学》（Heruistics and Biases: The Psychology of Intuitive Judgment, 2002），它在这一方面提供了比较完整而详细的描述。卡尼曼的《思考：快与慢》（Thinking, Fast and Slow, 2011）包含了对这方面的理念的指导（对没有受过任何数学训练的初学者尤其适用）。此外，还有哈斯戴（Hastie）和达维（Dawe）的《不确定世界的理性选择》（Rational Choice in an Uncertain World, 2001）和拜农（Baron）的《思考和抉择》（Thinking and Deciding, 2008）以及尼克尔森（Nickersn）的《认知和几率：概率推理的心理学》（Cognition and Chance: The Psychology of Probabilistic Reasoning, 2004）。

本章中所讨论的概率思维具有重大的实践意义。由于没有充分运用概率思维能力，医生们选择了效果欠佳的治疗方法（Groopman, 2007）；人们不能准确地评估环境风险（Gardner, 2008）；在法律程序中错误地使用信息（Gigerenzer, 2002; Gigerenzer et al., 2007）；动物不断被捕杀，以至濒临灭绝（Baron, 1998）；对病人实施了不必要的手术（Gigerenzer et al., 2007; Groopman, 2007）；有人做出了错误的财务判断，损失巨大（Zweig, 2008）。

当然，我们不可能在一个章节里全面地讨论统计推理。我们的目的就是想强调统计对于研究及理解心理学的重要性。不幸的是，当遇到统计信息时，我们还找不到一个放之四海皆准的规则。程式化的推理技能不像科学思维中的其他部分那么容易获得，而是需要通过正规学习才能掌握。

尽管很多科学家都真诚地希望一般大众能够知悉和理解科学知识，但有时对一门学科的精通依赖于对某些信息的掌握，而对这些信息的掌握又只有通过正规的学习才能实现。如果说对一门学科的深入理解是一般外行人也能随便达到的，那是一种在学术上不负责任的态度。统计学和心理学就属于这一类学科。不精通统计和概率的人不可能成为合格的心理学家（Evans, 2005）。美国心理科学协会主席莫顿·安·格恩斯巴彻（Motron Ann Gernsbacher, 2007）从智力价值中选出了10项，她认为这10项是通过心理学训练逐渐灌输的。10项中的4项都是数据和方法学领域的。极具威望的APA教学奖获得者鲁迪·班杰明（Ludy Banjamin）论述了在心理学导入课堂上应该说的内

容。虽然承认在这样的课堂上必须呈现学科上最重要的发现，班杰明认为：“从长远来看，教学生评估数据是同等重要的。在六周后的测试中，他们不会记住负强化和惩罚的区别，但是如果他们记住在课上讲过的对数据的批判性思维……那是我真正希望看到的教学遗产。”（Dingfelder, 2007, p.26）

我们当前的世界被数据和显示数字的图表充斥着。在医药、金融、广告、新闻行业，要求我们有统计的基础（Lutsky, 2006），我们需要学习对它们进行评估。幸运的是，学习心理学对具备统计直觉和洞察力有独特的作用。作家纳塔利·安吉尔（2007）在其很受欢迎的一本有关科学的重要发现的书中明确指出，概率和统计在诸多学科中处于核心地位。该书覆盖了所有的学科，但是在他书中开头部分，确切地说是在第2章，安吉尔介绍了理解概率和统计的重要性。

不可否认，本书的一个目的就是要使心理学的研究能为广大读者所接受。然而，心理学进行理论建构所依靠的实证方法和技术与统计学是如此密不可分（这一点和其他很多领域一样，如经济学、社会学和遗传学），以至于没有一个人可以在对统计学毫无知晓的情况下精通心理学。因此，尽管这一章对于统计思维介绍得相当粗略，但它的主要目的是要凸显另外一个对于理解心理学至关重要的专业领域。

小结

和大多数学科一样，心理学研究所得出的是概率式的结论——大多数情况下会发生，但并非任何情况下都发生。虽然这些结论并非是百分之百地准确（就像其他科学中的情况一样），但根据心理学研究及理论所作出的预测仍然是有用的。阻碍人们理解心理学研究的一个原因就是，人们很难用概率的术语来思考。在这一章里，我们讨论了几个相当精彩的研究实例，这些例子表明，大多数人如何与概率推理背道而驰：当人们遇到具体的、具有鲜活性的证据时，就把概率信息抛到一边了。他们没有考虑到，较大的样本能够提供对于总体数值更为精确的估计。最后，人们表现出赌徒谬误（把原本无关的事件看成是有联系的）。赌徒谬误源于第11章将要讨论的一个更为普遍的倾向：未能认识到偶然性在决定结果时所起的作用。

Chapter 11
偶然性在心理学中扮演的角色

在第10章里，我们讨论了概率趋势、概率思维和统计推理的重要性。本章将沿袭这一话题，重点强调人们理解随机性和偶然性这两个概念时所遇到的问题。我们将强调，由于没有领会偶然性是如何始终贯穿于心理学理论中的，人们常常误解了研究对于临床实践的贡献。

试图解释偶然性事件的倾向

我们大脑的进化始终以这样一种方式，就是让我们能够不懈地寻求世界中的各种模式。我们寻求身边事物的关系、解释及其背后的意义。埃里克·瓦戈（Eric Wargo, 2008）在美国《心理协会观察者》（APS Observer）上写道："脑可以被描述为一个'无由来的关联性器官'——贪得无厌的意义制造者。"（p.19）

然而，这种极具生存适应性的人类认知过程有时也会反戈一击。例如，环境中没有什么可以进行概念化的东西，可我们还是一味地去寻求概念性的理解，这就是一种不良适应。那么，到底是什么在人类认知这一最与众不同的方面制造麻烦呢？是什么打乱了我们对结构的寻求并阻碍了我们对事物的理解呢？你猜对了，是概率。说得更具体些，是偶然性和随机性。

偶然性和随机性是我们周围环境不可分割的一部分。偶然性和随机性的规律支配着生物进化和基因重组的机制，物理学也运用关于偶然性的统计定律来解释物质的基本结构。自然界发生的很多事情都是系统性以及可解释的因素与偶然因素共同作用的结果。再回想一下前面谈到的例子：吸烟导致肺癌。生物学上系统的、可解释的方面将吸烟和某一疾病联系起来，但这并不表示所有吸烟者都会患肺癌，这种趋势是概率性的。或许最终我们能解释为什么有些吸烟者不会患肺癌，但在现阶段，这种变异性必须归因于大量偶然性因素，是这些因素决定着一个人是否患某一疾病。

这个例子说明，当一件事取决于偶然性时，并不一定表示它是不确定的，只是说它目前是无法确定的。掷硬币是偶然事件，但并不是说在对抛掷的角度、硬币内的金属成分以及许多其他变量加以测量之后，也不可能确定其抛掷的结果。实际上，这些变量确实决定了掷硬币的结果。但是，我们称掷硬币为随机事件，是因为在每一次抛掷时，我们没有比较简易快捷的方法来测量这些变量。一次抛掷的结果并不是严格意义上的不确定，它只是在当下无法确定而已。

世界上的许多事件不能以系统性的因素来完全解释清楚，至少现在还不能。然而，当一个特定的现象没有现成的系统解释的时候，我

们头脑中的概念寻求“设备”往往仍在隆隆运转，试图将无意义的理论强加于原本随机的数据。心理学家曾对此现象进行了实验研究。在一个实验情境中，要求被试观察一系列在多个维度上有所区别的刺激物，并告诉他们其中的一些刺激物属于一类，而其他的则属于另一类，被试的任务是去判断每一个刺激物属于这两类中的哪一类。实际上，刺激物是研究者随机归类的，因此除了随机性，并没有任何其他规律。但是，被试很少敢做随机猜测。相反，他们通常会煞费苦心地编织出一套理论来解释刺激是如何分配的。

许多金融分析师的思维方式证明，认识随机性对于特定领域的巨大影响是多么困难。金融分析师通常会对股票市场价格的每一次小的波动都编造出精细的解释，而实际上这种变化大多只是随机波动而已（Kahneman, 2011; Taleb, 2007）。我们每晚应该在电视上听到的是“由于存在交互作用的系统的随机波动，道琼斯今日平均上涨27个百分点”。但你永远听不到这种头条，因为金融分析师会暗示他们能够解释一切——交易行为的每一次小的波动。他们不断地对客户暗示他们可以（也许他们也相信自己可以）“征服市场”，即使当大量的证据表明他们中的大部分其实是做不到这一点的。过去几十年中，如果你购买了标准普尔指数中的所有500种股票，然后放着不去管它（我们称之为“傻子策略”的办法——去买一种依照这一指数的互惠基金），那么今天你获得的回报会比2/3的华尔街股票经纪人为他们的顾客所赚得还要高（Bogle, 2010; Malkiel, 2011; Mamudi, 2009; Regnier, 2010），你的成绩也会打败80%订阅费已经涨至每年1000美元的财经通讯杂志。

但是，我们要如何看待那些确实战胜了傻子策略的经纪人呢？你可能想知道这是否意味着他们具有某些特殊的才能。我们通过设想这样一个实验来回答这个问题：有100只猴子，每只猴子手中握有10支飞镖，它们都向一面写有标准普尔500指数的墙上掷飞镖，飞镖扎中的股票就是那年要买的股票。那么，一年后它们的业绩会是怎么样的呢？有多少只猴子能打败标准普尔500指数？恭喜你答对了，大概有一半的猴子会。那么，你会不会愿意付钱给这一半打败标准普尔500指数的猴子，授权它们在下一年帮你选股呢？

这个关于财经预测的例子的延伸，证明了原本纯粹随机的事件会因怎样的逻辑而看起来像是由可以预测的因素造成的（Paulos, 2001）。假想你收到一封信，信中告诉你有这样一份关于股票市场预测的信息。这个信息并不收费，只是要求你试试照着他们的建议去买股票，然后看看它的预测灵不灵。它告诉你IBM股票会在下个月攀升。你把这份资料随手一扔，但是你确实注意到在下一个月里IBM股

票果真涨了。如果你曾读过一本与本书的内容类似的书，你会觉得这是稀松平常的事情，仅会将其视为一次侥幸的猜中。后来，你又收到另一份来自同一家投资咨询公司的信息，该信息说IBM股票会在下个月下跌，当股票确实跌了的时候，你仍将其视为侥幸，但是这一次你可能就有点儿好奇了。当这家公司寄来第三份资料，预测IBM下个月会再次下跌时，你发现自己对这几页财经内容的关注度提高了。继而你发现该信息又一次作出了准确预测，IBM这个月确实又下跌了。当来自这家公司的第四份资料说IBM下月会涨，而且也确实涨了时，你难免会觉得这个信息真还挺神，从而情不自禁地想花29.95美元去订一年这本如此有价值的资料。这种诱惑难以抵挡，除非你能想象这样的场景：此时在一个简陋的地下室里，某人正在准备下周要寄出的1600份资料，这些资料会按电话黄页上的1600个地址发出，其中800份预测IBM下月上涨，800份预测下跌。当IBM在下个月真的涨了，公司就继续把资料只发给上月接收到正确预测的800位“客户”（当然，其中还是400份预测涨，另外400份预测跌）。然后，你可以想象，这个“锅炉房”——可能还包括在背后煽风点火、辅助造势的电话营销骗子，正在向第二周接收到正确预测的400位客户发送第三个月的预测信息（还是200份预测涨，200份预测跌）。是的，你就是连续四次收到正确的随机预测信息的100个幸运儿之一！这100个“幸运儿”中的大多数会为了能继续收到信息而支付29.95美元。

现在看来，这就像是一个玩弄众人于股掌之上的可怕骗局。实际也是如此。而当那些“受人尊敬”的财经杂志或电视节目给你推荐“连续四年击败一半以上对手的股票经纪人”时，情况也好不到哪儿去。请回想一下猴子掷飞镖的场景，设想这些猴子是年年选股的股票经纪人。很明显，第一年他们之中有50%会击败他们的对手；第二年，这50%的人中又有一半——按随机水平来说——会击败其对手，即25%的经纪人能连续两年击败他们的对手。之后，第三年又有一半——随机水平——能击败对手，即总人数的12.5%连续三年击败对手。最终，到第四年，又会有这些人的一半（总人数的6.25%）能击败自己的对手。因此，100只猴子中大概有6只能取得像财经节目和报纸所说的“连续四年击败了其他的经纪人”骄人成绩的猴子。那么，这6只击败了一起掷飞镖的同伴的猴子（正如你所见，也击败了大多数现实生活中的华尔街经纪人；参见Egan, 2005; Malkiel, 2008）的确有资格在电视节目“华尔街一周”中亮相，你觉得呢？

解释偶然性：错觉相关和控制错觉

人们有解释偶然事件的倾向，这一现象在心理学的研究中称为错

觉相关。当人们相信两个事件在通常情况下应该同时发生时，就会认为自己频繁地看到了同时发生的现象，甚至当这两个事件的同时出现是随机的，并不比任何其他两个事件同时发生的频率更高时也是如此。简言之，即使是面对随机事件，人们也倾向于看到他们所期望的关联。他们在原本没有结构的地方看到了结构（Kahneman, 2011; Whitson & Galinsky, 2008）。

许多有控制的研究都证明，当人们头脑中已经预设了两个变量相互关联的想法时，他们甚至能够在两个变量根本毫无关系的数据中发现联系。不幸的是，这一发现在现实生活中也广泛存在，并对人们的生活产生负面影响。例如，许多从事心理治疗工作的人一直都对罗夏墨迹测验的效度深信不疑。这个著名的墨迹测验要求被试对一张白纸上的墨迹作出反应。因为这一墨迹缺乏结构，所以其理论是人们会以自己对模糊情境的典型反应来对这些墨迹作出反应，从而揭示其“潜藏的”心理特质。这种测验也被称为投射测验，因为它假定被试会将他们潜意识的内心活动和感受投射到墨迹上。然而，问题是没有任何证据表明当罗夏测验作为一个投射测验而使用时，提供了任何额外的诊断价值（Lilienfeld et al., 2010; Wood, Nezworski, Lilienfeld, & Garb, 2003）。对罗夏测验的信心是源自于错觉相关现象。临床心理医生从病人的反应模式中看到了关联，是因为他们相信本来就有这种关联，而不是真的从反应模式中观察到了什么关联。

在我们的生活中，许多人际交往里都包含大量的偶然成分：“双盲约会[1]最终促成了婚姻；取消约谈而丢了工作；误了班车而遇到了高中老同学，等等。认为生活中每一件偶然的小事都需要精细的解释，这种思维固然不对。但是，当偶然事件确实会产生重要的后果时，人们不免要建构一些复杂的理论去解释它们。

试图去解释偶然事件的倾向可能源于我们深切地渴望相信自己是可以控制这些事件的。心理学家已经对所谓“控制错觉”（illusion of control）现象进行了研究，这一现象指的是人们有一种倾向，愿意相信个人能力可以影响偶然事件的结果（Matute, Yarritu, & Vadillo, 2011）。这一错觉广泛存在的证据来自于美国各州彩票发行的经验。这些州充斥着教人们如何“征服”彩票的伪科学书籍。这类书之所以畅销，是因为人们不懂得随机性的含义。事实上，自从20世纪70年代中期新泽西州引入了参与式彩票售卖之后，美国各州才爆发购买彩票的热潮。所谓参与式就是让购买者可以自行刮奖或自行挑选号码。而这类参与性博彩正是利用了当时兰格研究的控制错觉现象：人们错误地相信他们的参与行为能够决定随机事件。这种错觉在某些喜欢赌博的人身上非常强烈，他们愿意花1495美元学习被认为能够帮助他们控制

掷骰子结果的所谓“特殊课程”（Schwartz, 2008）。这样的“课程”当然是彻头彻尾的骗局。

还有一些心理学家则研究了另一个与此相关的现象，该现象被称为“公平世界假设”。它是指人们倾向于相信自己是生活在一个公平的世界里，在这里，每个人都得到他们应得的东西（Hafer & Begue, 2005）。研究者发现了一些实验证据，证明了公平世界中存在一种“罪有应得”的信念：人们会鄙视那些偶然不幸的受害者。为偶然事件寻求解释的倾向导致了这一现象。人们很难相信一个完美无瑕的或是道德修养高的人会因为偶然事件而遭遇不测。固然我们想要相信“好人有好报、恶人有恶报”，但是偶然性是不偏不倚的，它以完全不同的方式运行：好事坏事都以相同的概率发生在不同人身上，它不会对“好人”有所眷顾。

公平世界假设中所体现的对于偶然性的错误理解，也助长了其他一些错误的民间信念，导致人们容易看到虚假相关。例如，我们在第6章中提到过，“盲人有非常敏锐的听觉”就是一个错误的信念，这个错误信念可能会一直流传下去，因为这种联系能体现“上天是公平的”，而这正是人们希望看到的。

注释

[1] blind date，即安排互不相识的男女进行约会。——译者注

偶然性和心理学

在心理学中也存在这样的倾向：研究者试图解释一切，希望其理论不仅能解释行为中系统的、非随机的成分，还要能解释任何细微的变异。这种倾向导致了不可证伪的心理学理论的泛滥，既包括个人提出的理论，也包括那些看似科学的理论。“心理历史学”的奉行者常常犯下此类错误。一个著名人物生命中的每一个细小的变化及转折，都经由精神分析学派的理论在心理历史中得以诠释。大多数心理历史事件存在的问题是，不是它们解释得太少，而是它们解释得太多。这一研究方法的奉行者很少承认一个人的一生是由许多偶然因素决定的。

对于想要运用心理学知识的外行人来说，理解偶然性这一因素的作用是非常重要的。受过正规训练的心理学家承认他们的理论只能解释人类行为变化的一部分而非全部，他们会坦然面对偶然因素。但是，那个在奥普拉脱口秀中出现的（见第4章开头）能对每一个个案及人类行为的每个细节作出解释的嘉宾，引发的不是崇拜而是质疑。真正的科学家从不惧怕承认自己的无知。总之，评价心理学主张的另一实用法则就是：在接受对某个事件的复杂解释之前，先想一想偶然因素在其中扮演了什么角色。

巧合

为纯粹偶然的事件寻求解释的这种倾向，也导致我们对许多巧合事件的性质产生误解。许多人认为巧合需要特别的解释，他们不理解巧合的发生并不需要偶然性之外的因素，巧合并不需要特别的解释。

大多数辞典将“巧合”定义为“有关联的事件意外地、不可思议地同时出现了”。鉴于这本字典把意外定义为“偶然地出现”，所以这个定义不存在问题。巧合只是相关事件偶然地同时出现。不幸的是，许多人并不这样解释巧合。那些在事件中寻求模式和意义的倾向与巧合“不可思议”的特性结合在一起，让许多人忘记他们可以用偶然这一因素来解释巧合，反而为理解这一现象寻求特别的解释。下面讲的这个故事你一定已经听过无数次了：“那天我正坐在那儿寻思，我好久没给

得克萨斯州的老比尔叔叔打电话了，紧接着电话铃就响了，你猜怎么着！正是我那老比尔叔叔打来的。这种心灵感应的背后肯定有点儿什么原因！”这就是一个典型的为巧合事件编造解释的例子。每天，我们大多数人都可能想到很多或远或近的人，这些人在我们想起他们时，有多少人可能会打电话来呢？几乎没有可能。这样一年之内，我们可能想过数百个不曾打来电话的人。最终，在经历数百次这种我们不曾意识到的“错误尝试”之后，某个人在我们想他/她的时候正准备给我们打电话。这种事情难得一见，但难得一见的事情也会发生——纯粹是偶然。其他解释都是画蛇添足。

如果人们真正理解了巧合的含义（一个偶然发生的令人不可思议的事件），他们就不会落入陷阱去寻求系统的、非偶然性的解释。但事实正相反，对很多人来说，巧合是需要偶然性以外的原因来解释的。例如，许多人都听到过这样的说法：“天哪！简直太巧了！我真想知道为什么！”这反映了一个基本的错误——巧合不需要解释。

心理学家大卫·马科斯（David Marks, 2001）建议大家今后用“罕见偶合”（oddmatch）这个比较中性的名词来形容令我们感到惊异的两个事件的同时出现。有一种错误信念助长了为巧合事件寻求解释的倾向，这种信念认为罕见的事不会发生，罕见偶合也绝非偶然。我们的这类错误信念之所以如此强烈，是因为概率有时是用几率（odds）[1] 这一词语来表述的，而这种表述具有双关的暗示作用。看看我们是用什么方式来表述概率的：“啊！天哪，这事儿是极不可能发生的！因为它出现的概率只有1/100！”我们在做这样的表述时所用的方式让人强烈地感觉到这件事绝不会发生。当然，我们可以用另外一种表达方式来表述同一件事，而这一方式可能给人带来完全不同的感受：“在100个同类事件中，这种结果可能会出现一次。”这种表述方式强调，尽管这一事件是少见的，但是从长时间来看，罕见的事终究一定会发生的。简而言之，罕见偶合是会偶然发生的。

事实上，概率法则确保了随着事件发生次数的增加，一些罕见偶合出现的可能性会变得很大。这一定律不仅允许罕见偶合出现，而且从长远来看几乎保证了它的出现。请看马科斯（Marks, 2001）的例子，如果一次掷5枚硬币，结果它们都是正面朝上，你将认为这是一个几率偶合，一件不太可能的事情。是的，它发生的概率是1/32或0.03。但是如果你将这5枚硬币掷100次，再问，在这100次中，至少有一次全部正面朝上的可能性是多少呢？答案是0.96, 就是说，100次中，这一罕见偶合是极有可能发生的。

简而言之，基本上你能想到的所有罕见偶合都会出现，只要你等待的时间足够长。1913年秋天，蒙特卡洛赌场的一盘幸运轮中，黑色

连续出现了26次（Kaplan & Kaplan）！

有一些网站关注“许多著名的音乐人都在27岁死亡”这一令人毛骨悚然的事实：艾米·怀恩豪斯（Amy Winehouse, 英国女歌手）、科特·柯本（Kurt Cobain, 涅槃乐队主唱，Grunge曲风代表人物）、吉姆·莫里森（Jim Morrison, 美国歌手，大门乐队主唱）、吉米·亨德里克斯（Jimi Hendrix, 美国民谣歌手，以吉他演奏技巧著称）、詹妮斯·乔普林（Janis Joplin, 美国民谣女歌手，嬉皮运动中的代表人物）等等（O'Connor, 2008）。只可惜根本就没什么可“毛骨悚然的”，也没什么好解释的，这就是随机事件。我们之所以知道是随机事件，是因为出版的英国医学杂志上刊登了对1956年到2007年之间1046个英国唱片史上曾经拥有榜首专辑的歌手进行的统计分析（Barnett, 2011）。分析结果显示，没有趋势让这些歌手一股脑地在27岁死亡。

懂得在什么时候避免对纯粹随机因素导致的事件编造复杂的解释，这是具有实际作用的。认知心理学家卡尼曼描述了在Yom Kipper战争中以色列空军打交道的事例。两个飞行中队出发并返航，一队损失了四架飞机，另一队则没有损失。军方希望卡尼曼调查一下，之所以有这样的差异是否有特别的因素在起作用。卡尼曼并没有去做调查，但是卡尼曼知道，以这样的小样本，任何找到的因素都有可能是虚假的——不过是纯粹的偶然性波动的结果而已。他没有去做调查，而是运用本章所谈到的理念去告诉以色列空军不要浪费时间：“我推论，运气是最可能的答案，对不显见原因的随机搜索其希望是渺茫的，同时遭遇损失的中队飞行员也不必因为觉得自己和战友有错而背上额外的负担。”（p. 116）。

个人的巧合

发生在我们个人生活中的罕见偶合往往对我们具有特殊的意义，我们尤其不愿将其归因为偶然。产生这种倾向的原因有很多，某些是动机性和情感性的，还有一些是概率推理的失败。我们通常不能意识到，罕见偶合只是巨大“概率事件”样本库中一个非常小的部分而已。对我们中的某些人来说，罕见偶合看起来好像经常发生，但是它真的经常发生吗？

想想，如果我们现在对你个人生活中的罕见偶合加以分析，会得到什么结果。假定某一天里你参与了100件不同的事情。考虑到现代工业社会中生活的复杂性，这个数字并没有高估，实际上可能还低估了。你看电视、打电话、与人面谈、讨论去工作或去商场的路线、做

烦人的家务、看书获取信息、在上班时完成复杂的任务，等等。所有这些事件都包含很多可单独记忆的成分。这样一算，100件事其实真不算多。不过，我们就按100件事情来算。罕见偶合是指其中两个事件不可思议地联系在一起了。那么典型的一天中这100件事之间共有多少不同的、两两匹配的组合呢？用一个简单的公式就能算出结果，你通常一天有4950个不同的配对组合，而一年有365天。

罕见偶合是令人难忘的，老比尔叔叔打来电话的那一天可能会让你数年难忘。假如你把10年内所记得的所有罕见偶合数出来，也许也就6或7件（或多或少，人们对于小概率有不同的标准）。这六七件事情来自一个多大的概率事件样本库呢？每天4950个配对事件，乘以一年365天，再乘以10年，得到18067500个配对。总之，10年中如果有6个你认为是罕见偶合的联系发生了，就有18067494个也可能是罕见偶合的其他配对事件发生了。所以，你的生活中的一个罕见偶合发生的概率是0.00000033。有6个罕见偶合出现在1800万个事件中，的确很稀罕，但并不奇怪。罕见的事件确实发生了，它们也的确少见，但是偶然性这一因素保证了它们一定会发生（回忆前面掷5枚硬币的例子）。在我们的例子中，6件奇事发生在你身上，它们可能是巧合：两个关联性事件由于偶然性的存在而不可思议地同时发生了。心理学家丹尼尔卡尼曼（2011）认为是我们的语言让我们力有不逮。我们有一些形容“先前的想法最终应验”的词语（预感、直觉），但是我们没有一类词语是标记并提醒“过去的信念被证明是错的”。大多数人不会自发地想到说“我预感婚姻不会持久，但是我错了”（p.202）。因为在某种程度上这对他们来说似乎很奇怪。没有标记这类情况的词语，我们也不会倾向于对先前的预测错误进行标记。

心理学家、统计学家以及其他科学家都指出，许多罕见偶合实际上并没有人们通常认为的那么“罕见”。著名的“生日问题”是最好的例子。在一个23人的班级里，有两个人生日是同一天的概率是多少？大多数人会认为非常低。而实际上，23人的班级中，两人同一天过生日的可能性大于50%。而在35人的班级，可能性就更大了（概率大于0.80; 见Martin, 1998）。因此，美国历史上有43位总统，詹姆斯·波尔克和沃伦·哈丁两位在同一天出生（11月2日）也就不足为奇了。同样地，有38位总统都已过世，其中米勒德·菲尔莫尔和威廉·塔夫脱死于同一天（3月8日）也不应令人感到惊讶，甚至还有另外3位总统——约翰·亚当斯、托马斯·杰菲逊、詹姆斯·门罗——都死于同一天，而这一天竟然是7月4日，美国独立日！后面这个神奇吗？不过是概率使然罢了。

注释

[1] 英语中“odds”既有“罕见”的含义，还有“几率”的含义。——译者注

接受错误以减少错误：临床预测与统计预测

在试图解释世界上发生的所有事，同时又拒绝承认偶然因素的作用，实际上会降低我们对现实世界的预测能力。在某个领域中，承认偶然因素的作用意味着研究者必须接受这样一个事实，即我们的预测不可能百分之百准确，预测中总是会犯一些错误。但有趣的是，承认我们的预测达不到百分之百的准确度，实际上反而有助于我们提高整体预测的精确性。这听起来好像有点儿矛盾，但是事实确是如此：为了减少错误，就必须接受错误（Einhorn, 1986）。

“我们必须接受错误以减少错误”这一概念可以通过一个在认知心理学实验室里研究了数十年的、非常简单的实验任务来证明。这个实验任务是这样的：被试坐在两盏灯（一红一蓝）前，实验者要求他们去预测每次测试时哪一盏灯会亮，被试要参与很多轮这样的测试，并按准确率给予一定的报酬。实际上，所有的测试都是在70%的次数亮红灯、30%的次数亮蓝灯的条件下进行的，两种灯以随机顺序出现。实验过程中，被试很快就感到红灯亮的次数比较多，因此也就在更多的测试中预测红灯会亮。事实上，他们确实在大约70%的测试中预测红灯会亮。然而，正如前面所讨论的，被试在实验过程中逐渐发现并相信灯亮是有一定模式的，但却从没想过序列是随机的。为了要使他们的预测百发百中，他们在红灯与蓝灯之间换来换去，保持70%的次数预测红灯会亮，30%预测蓝灯会亮。被试极少意识到，尽管蓝灯亮的概率为30%，如果他们停止在红灯和蓝灯之间换来换去，他们的预测会更好一些！为什么会是这样的呢？

让我们想想这一情境背后的逻辑。在以70：30的比例随机点亮红灯或蓝灯的情况下，如果被试在70%的测试中预测红灯会亮，30%的测试中预测蓝灯会亮，他的准确率会是多少呢？我们将用实验中间部分的100个测试来计算——因为那时被试已经注意到红灯亮的次数比蓝灯多，从而开始在70%的测试中预测红灯会亮了。在100次测试中有70次红灯亮了，所以被试在这70次中有70%的正确率（因为被试在70%的测试中预测红灯会亮），也就是说，被试在70次中有49次正确的预测；100次测试中有30次蓝灯亮了，被试在这30次中有30%的正确率（因为被试在30%的测试中预测蓝灯会亮），也就是说，被试在30次中有9次正确的预测。因而，在100次测试中，被试的正确预测

是58次。但是请注意，这是多么可怜的成绩啊！如果被试在注意到哪一盏灯亮得比较多后，就总是预测那盏灯会亮——在本实验中，就是注意到红灯亮的次数比较多，因此就总是预测红灯会亮（姑且称之为“百分百红灯策略”），那么，他在100次测试中会有70次正确的预测。虽然在蓝灯亮的30次测试里，被试将没有一次正确的预测，但是总准确率仍然高达70%——比在红灯与蓝灯之间来回变换以追求“百发百中”的58%的准确率要高12个百分点！

最优策略也意味颇多——每次蓝灯亮起你都错了。而且，由于蓝灯总会亮若干次，永不押蓝灯似乎也不对，但这却是正确的概率思维所需要的，它要求接受在蓝灯上所犯的错误，以换得每次都押红灯之后整体命中率的提高。以一定的精度预测人类的行为时，有时也需要接受错误以减少错误，也就是说，在依靠一般性的原则来做出比较准确的预测的同时，也要承认我们不可能在每件具体事情上都对。

但是，“接受错误以减少错误”做起来很难。在心理学领域里，40年来关于临床预测和统计预测的研究就证明了这一点。统计预测是指依据统计资料中得出的群体趋势所作的预测。本章一开始所讨论的群体（也就是总体）预测就是属于这种预测。一种简单的统计预测是，针对凡是具有某种特征的所有个体，作出相同的预测。举个形象的例子，预测不吸烟者的寿命是77.5岁，而吸烟的人是64.3岁，就是一个统计预测。如果考虑的群体特征不止一个（运用第5章谈到的复杂相关技术——尤其是多元回归技术）将令我们的预测更加准确。例如，预测吸烟、肥胖且不运动者的寿命是58.2岁，就是在一个多变量（吸烟行为、体重和运动量）基础上的统计预测，这样的预测总是比单变量的预测更加准确。

统计预测在经济学、人力资源、犯罪学、商业与市场学以及医学等领域都很常见。例如，发表在《美国医学协会杂志》（the Journal of the American Medical Association）和《内科医学年鉴》（Annals of Internal Medecine）上的研究中报告了如下的概率趋势：在中年时期，肥胖的人出现心脏问题的概率比65岁以上不肥胖的人高四倍；与之相似，超重（但不肥胖）的人出现肾脏问题的概率是常人的两倍；而肥胖的人出现肾脏问题的概率是七倍（Seppa, 2006）。但是概率预测承认错误。不是所有肥胖的人都会有健康问题。回想政治播音员蒂姆·拉瑟特的例子（第10章），他58岁死于心脏病，医生判断拉瑟特先生10年内死于心脏病的概率只有5%。这意味着大多数（100人中的95人）和拉瑟特情况相似的人10年内不会患心脏病。拉瑟特先生就是那不幸的5%——他是一般趋势的例外。

然而，人们有时发现，根据精确证据采取行动很困难，因为这样

做需要心理上的训练。例如，2003年美国食品和药品管理局举行健康咨询会给出一条健康建议，警告说流行的抗抑郁药和青少年自杀之间存在潜在关联。很多医生担心，在统计基础上，这一警告会导致更多自杀事件。他们担心因药物引发自杀的青少年会减少，但会有更多的青少年死于犹豫是否开药。事实上，这也真的发生了。用这些药物治疗给孩子们带来暂时的风险，但不治疗抑郁情况会更严重。多数医生认为，这一警告付出的人命代价比拯救的多（Dokoupil, 2007）。这就是为这种情况所算的一笔账。或者我们可以说：这是精确预测的计算。但是听从一些世俗观念很难用得失计算，例如“安全比遗憾更好”。而在医学治疗中，“安全比遗憾更好”忽略了与之对等的另一半。他将我们的注意焦点放在了那些可能被治疗伤害的人身上，但是完全忽略那些因接受不到治疗而受到伤害的人。

在心理学的许多分支领域，如认知心理学、发展心理学、组织心理学、人格心理学与社会心理学中，其知识都是通过统计预测来表述的。相反，一些临床心理从业者则声称他们可以超越群体预测，对特定个体作出百分之百准确的预测，这种预测被称为临床预测或个案预测。与统计预测相反，临床预测是这样的：

> 职业心理学家声称，他们能对个体进行预测，从而超越了对“一般人”或不同类别的人所进行的预测……某些心理学家最大的不同在于，他们主张将每个人理解为独一无二的个体而不是群体的一部分，而统计概括是适用于群体的。某些心理学家声称能分析出在个体的生活中“什么导致了什么”，而不说“总体而言”什么是对的。（Dawes, 1994, pp.79-80）

临床预测似乎可以视为对统计预测的有益补充，但问题是临床预测并不准确。

如果证明临床预测是有效的，那么一个临床医生与他的病人接触的经验以及有效运用病人所提供的信息，应该使他能够提出比较好的预测。这个预测一定能胜过对病人信息进行编码，然后输入能够对量化数据加工的统计程序所得到的预测结果。总之，有人主张说，临床心理从业者的经验使得他们能够超越尚未由研究揭示的关系。“临床预测是有效的”这一观点很容易验证，不幸的是，经过检验，这一观点被证明是错误的。

对临床预测与统计预测的比较研究所得到的结果始终是一致的。自从保罗·米尔（Paul Meehl）的经典著作《临床预测与统计预测》（Clinical Versus Statistical Prediction）于1954年出版以来，60年间有超过100个研究表明，在几乎每一个曾经验证过的临床预测领域

（精神治疗的效果、假释行为、大学生毕业比例、电击治疗的反应、累犯问题、精神病住院治疗期的长短，等等），统计预测都优于临床预测（Kahneman, 2011; Morera & Dawes, 2006; Swets et al., 2000; Tetlock, 2005）。

在多个临床领域中，研究者给临床心理医生一份病人的信息，让其预测这个病人的行为。与此同时，他们也把同样的信息加以量化，用一个统计方程加以分析，这一方程是以先前研究发现的统计关系为基础编制的。结果都是统计方程大获全胜。这就表明，统计预测比临床预测更为准确。事实上，即使是在临床心理医生可以获得比统计方法更多的资料的情况下，后者仍然比前者的预测更准确。也就是说，临床心理医生除了拥有与统计预测一样的量化资料以外，还拥有与病人单独接触和访谈所得到的资料，但是这并没有令其预测变得像统计预测那样准确。即使拥有信息优势，临床判断仍然不能超越统计方法。产生这种结果的原因当然是统计方程将各种信息数据按照优化标准整合起来，并且做得准确而稳定。稳定这个因素能够让临床心理医生通过非正式方法收集到的资料和信息的优势消失殆尽。

在检验临床 - 统计预测的研究文献中，还包含这么一种方法，那就是给临床心理医生由统计方程得来的预测结果，让其根据自己与病人接触的经验来对这一预测作出调整。结果，临床医生对统计预测作出调整后，预测的准确度非但没有增加，反而降低了（见Dawes, 1994）。在这里，我们又看到了一个不能“接受错误以减少错误”的绝好例子，与前面所述的那个红蓝灯预测实验非常类似。应当利用灯亮次数多少这一统计信息而采用每次都预测红灯的策略（可以获得70%的正确率）时，被试却为追求次次正确而在红灯与蓝灯之间换来换去，结果正确率反而降低了12%（只有58%的次数是正确的）。同样地，在上述研究中，临床心理医生相信，他们的经验应该可以提供给自己一些“洞察力”，从而得以作出比定量数据更好的预测。实际上，这些“洞察力”根本不存在，他们的预测比依赖公开的统计信息所作出的预测要差。最后需要指出的是，统计预测的优越性并不局限于心理学，它业已扩展到了许多其他临床科学中——例如，医学（Groopman, 2007）、金融服务（Bogel, 2010; Kahneman, 2011）以及体育训练（Moskowitz & Wertheim, 2011）。

对于研究显示统计预测相对于临床预测的优势，保罗·米尔（Meehl, 1986）曾说：“社会科学中，没有任何一个争议能如此这般从这么大量的、性质上如此多样的研究中得到如此一致的结论。”（pp. 373—374）但令人尴尬的是，心理学领域并没有应用这一知识。例如，这个学科在研究生入学与心理健康培训招生等程序中

仍然不停地使用个人面试，尽管大量证据表明，面试方法缺乏效度。临床工作者也继续利用一些似是而非的证据来证明他们对于“临床直觉”的依赖是合理的，而不依靠更有效的总体性预测。例如，道斯等（Dawes et al., 1989）曾指出：

> 一种普遍的反统计论调或误区在于，认为群体统计不适用单个人或事。这种观点是对概率基本原则的误用……要保持逻辑上的一致，反统计论的鼓吹者就必须相信并承认，如果一个人被迫玩一次俄罗斯轮盘赌，允许他选择膛内装有1发或5发子弹。事件的单一性使得选哪把枪都无所谓（p. 1672）。

关于这一点的一个类比是，问问你自己对以下的科学发现有什么样的反应，这个发现是：完成过多次类似手术的医生，在下一例手术中成功的概率会比较高（Grady, 2009; Groopman, 2007）。现在医生A常做某一类手术，失败的可能性很小，而医生B从没做过这种手术，失败率可能很高。请问，你愿意让这两个医生中的哪一个来为你做手术呢？如果你相信“概率不适用于个案”，那你就不该介意让医生B给你做手术。

在诸如心理治疗效果等问题上，承认统计预测优于临床预测并不会对心理学的声望造成任何损失，因为在医学、商学、犯罪学、会计学甚至是家畜鉴定等许多领域中，这条规律都适用。尽管从总体上说，心理学不会因为这些研究结果而有什么损失，但是对那些以“专家”身份出入各种活动，并让病人相信他们有独一无二的临床个案知识的临床心理从业者来说，当然会造成声誉或者收入上的损失。

实际上，如果我们将“接受错误以减少错误”变为一种习惯，心理学和整个社会都将从中受益。在试图对每一个不同寻常的事件作出独特解释时（就我们目前的知识情况来说，独特的解释也许根本不可能），我们常常丧失了对更多平常事件的预测能力。请大家再次回想一下红灯—蓝灯实验，诚然，“百分百红灯策略”会对出现概率较小或很少出现的不寻常事件（蓝灯亮）作出错误的预测，但如果我们把注意力放在出现概率较小的事件上，采用“70%红灯、30%蓝灯策略”，结果会怎样呢？我们会在30个不寻常事件中正确预测9次（30×0.3），其代价是丧失了对21个常见事件作出正确预测的机会，没有对红灯作出70次的正确预测，只获得49次的正确预测（70×0.70）。临床领域中的行为预测也遵循相同的逻辑，为每一个案编造复杂的解释，确实可能抓住一小部分不寻常事件——但这是以损失了对大多数事件的正确预测为代价的，而在此方面，简单的统计预测则更有效。

瓦格纳和科仑（Wagenaar & Keren, 1986）论证了对个人知识的过分自信以及对统计信息的忽视会破坏“系安全带驾车”的交通安全推广活动的效果。因为人们总是认为：“我和别人不一样，我驾车很安全”。问题是，大多数人都认为“自己的技术比一般驾车者高明”（De Craen, Twisk, Hagenzieker, Elffers, & Brookhuis, 2011）——这显然是很荒谬的。

“统计数据不适用于单一个案”这一同样的谬误，是导致赌徒积习难改的重要因素。瓦格纳（1988）在他的赌博行为研究中总结道：

> 从我们和赌徒的讨论中可以非常清楚地看出，赌徒大体上都能意识到赌博造成的不良后果。他们也知道最终输的会比赢的多，而且在未来也是如此。但他们却不能把这些统计性的思路应用到下一局、下一小时或下一个晚上。丰富的直觉经验还是让他们觉得，统计学在下一局或下一小时里派不上用场，他们相信自己能够预测下一局的结果（p. 117）。

瓦格纳发现，无节制型赌徒对“接受错误以减少错误”有很强的排斥倾向。例如，21点牌局的玩家，普遍拒绝使用一种基本策略，这种基本策略可以保证把庄家的胜率从6%或8%降低到不足1%。基本策略是一个长期性的统计策略，无节制型赌徒之所以拒绝它，是因为他们坚信“有效的策略应该是在每一把都有效”（p.110）。瓦格纳研究中的赌徒“总是一成不变地说，这类系统的一般性策略是不会奏效的，因为它们忽略了每一个具体情境的独特性”（p. 110）。这些赌徒抛弃能保证他们少输上千美元的统计策略不用，转而去徒劳地追求建立在每一具体情境独特性基础之上的“临床预测”。

另一个统计预测经常打败临床预测的地方是体育领域。许多人在2011年观看了改编自迈克尔·刘易斯（Michael Lewis）原著的电影《点球成金》（Moneyball）。故事讲述了奥克兰竞技队经理比利·宾（Billy Beane）的故事，宾否决了他的球探的“临床”判断（过于依赖可见的身体条件），依靠球员过往表现的统计资料去评估准备招入的球员。他的球队表现得物超所值，他从棒球统计学家那里借鉴来的精算法随后被许多团队复制。在其他运动领域，统计方法比“教练的判断”更有优越性（见 Moskowitz & Wertheim, 2011）。

当然，这里有关临床—统计预测研究文献的讨论，并不意味着个案研究在心理学中毫无价值。请大家记住，这一章所谈的只是“对行为的预测”这一特定情境。回想一下在第4章中对于个案研究价值的讨论，个案信息在引发对重要的、需要进一步研究的变量的关注方面是非常有用的，而这一章中所说的则是一旦相关的变量已经确定，我们

要开始运用它们来预测行为时，测量这些变量并使用统计公式来进行预测始终是最优程序：首先，我们通过统计方法得到了更为准确的预测；其次，统计方式优于临床预测之处在于，统计程序所得出的预测是公共知识，任何人都可以使用、修改、批评或争论。相反，如果使用临床预测就等于要依靠个别权威的评估——由于这类判断太过个别和特殊，因此不能接受公众的评议。

小结

偶然性在心理学中扮演的角色时常被外行人士和临床心理从业者所误解。人们很难认识到，行为事件结果的变化中有一部分是由偶然因素造成的。也就是说，行为的变化有一部分是随机因素作用的结果，因此心理学家不应自诩能够预测每一例个案的行为。心理学的预测应该是概率性的——是对总体趋势的概率性预测。

认为自己可以在个体层次上进行心理预测，是临床心理学家常犯的错误。他们有时候会错误地暗示别人，临床训练赋予了他们一种对个别案例作出准确预测的“直觉”能力。恰恰相反，几十年来，有价值的研究都一致表明，在解释人类行为的原因方面，统计预测（基于群体统计趋势的预测）远远优于临床预测。目前还没有证据表明，临床直觉能预测一个统计趋势是否会在一个特定的个案身上出现。因此，当对行为进行预测时，千万不要对统计资料置之不理。统计预测也昭示，当对人类的行为进行预测时，错误和不确定性将始终存在。

Chapter 12 不招人待见的心理学

罗德尼·丹杰费尔德（Rodney Dangerfield）是风靡30多年的一位美国喜剧演员，他标志性的口头禅是：“我得不到尊重！”从某种意义上说，这也正是心理学在一般公众心目中的形象写照。本章就是想谈谈为什么心理学会像丹杰费尔德那样无法获得应有的尊重。

虽然公众对心理学话题怀有浓厚兴趣，但他们对于心理学及其所取得的成就给出许多负面评价。心理学家们都意识到了这个“形象问题”，但他们又感到无能为力，所以干脆不去管它，这样做其实是错误的。当大众传媒在决定公众感知（例如，虚构的电视“纪录片”对那些知识储备不足的公众来说就成了真实的历史）方面越来越有影响力的时候，不理会心理学的形象问题只会让情况变得更糟。

心理学的形象问题

之前我们曾讨论过造成心理学形象问题的成因。例如，在第1章中所讨论过的弗洛伊德问题，无疑导致了人们对心理学较低的评价。如果要公众列举一个著名心理学家的话，这个人不是弗洛伊德，就是斯金纳（Overskeid, 2007）。弗洛伊德的精神分析在很多方面都是不符合科学的，但是就像第1章所描述的那样，这些不可证伪的想法在现代心理研究中没有任何作用。至于斯金纳，当一门学科中最具影响力的学者被误解为主张人没有思想、人和老鼠没有差别的时候，这门学科也就变得希望渺茫了。斯金纳当然没有否认人类能够思考（Gaynor, 2004），但是对他思想的歪曲版本太多，很少有人知道他从动物身上发现的许多有关操作性条件反射的定律已被证实的确能推广到人类行为上，但公众对这些科学事实知之甚少。

心理学和超验心理学

除了弗洛伊德和斯金纳的研究之外，外行人对其他卓越的心理学研究几乎一无所知。想证明这一点，到附近的书店去看看公众能买到什么样的心理学读物就知道了。你的调查会发现，那些摆在书店卖的心理学读物通常可以分为三类。第一类是心理学的一些早期经典著作（弗洛伊德、斯金纳、荣格、弗洛姆、埃里克森等），这些著作多半侧重老式的精神分析观点，已经完全不能代表当代心理学了。令心理学家感到沮丧的是，这一领域最有价值的著作都被淹没在书店的科学或生物学类书籍中。例如，心理学家史蒂芬·平克（Steven Pinker）的名作《思维的运作》（How the Mind Works）总是被归在科学类而非心理学类图书中。因此，他所探讨的关于认知科学的重要文章，被迫与生物学、神经生理学或计算机科学而非心理学为伍。

在多数书店中可以找到的第二类读物，是那些伪装成心理学的伪科学书籍，里面充斥着无数超自然现象，如心灵感应、千里眼、意念移物、超前感知、转世、生物节律、星灵投射、金字塔力量、植物沟通、通灵术等（Lilienfeld, Lohr, & Moirer, 2001）。书店里的心理类书籍中这类货色大量存在，无疑导致（也反映了）人们的误解：心理

学家就是证实这些超常现象存在的人。这种误解对心理学而言多少有点儿讽刺。事实上，心理学与那些超常现象之间的关系很容易说清楚。这些超常现象压根不在现代心理学感兴趣的范畴之内。个中缘由可能会令许多人大吃一惊。

超自然体验和其他超常能力的研究不被认为是心理学的一部分，此观点可能会激怒许多读者。多个调查结果都显示，超过40%的公众相信超感知现象的存在，并且狂热地信奉自己的信念（Farha & Steward, 2006; Kida, 2006; Shermer, 2011）。历史研究和调查指出了公众热衷于这类信念的原因（Begley, 2008; Humphrey, 1996; Park, 2008; Stanovich, 2004）。如同大多数宗教一样，许多所谓的超常现象也标榜诸如转世之类的说法。对部分人来说，来世的说法能满足其超越现有生命极限的需求。

心理学研究“不识时务”地指出超感知得不到证实，无疑粉碎了这些人的热切企盼。心理学主张不把超感知视为一个可行的研究领域，不可避免地会引发其信徒的不满，他们控诉说，心理学家把这一类主题排除于心理学研究之外的做法是独断专行的。如果心理学家仅是摆出无可奈何的姿态并无视这些反对的声音，这将无助于增进公众的理解。与之相反，心理学家应该针对这些反对观点的谬误给予细致而清晰的解释。这样的解释要强调科学家们并不是根据什么法令来确定研究主题的，也没有什么条例指出什么能研究、什么不能研究。研究领域的兴起、延续或终结，所依据的是理论及方法的自然选择过程。那些产生出丰硕理论及实证发现的领域都获得了大量科学家的认可；而那些理论上行不通或者没有能被重复验证的领域就会被摒弃。这种对理论与方法的自然选择引导着科学向真理靠近。

例如，超感官知觉在当代心理学中不被认为是一个可行的研究主题，就是因为其研究一直无法积累任何正向的成果，所以它让大部分心理学家失去了兴趣。在这里，我要强调“当代”一词，是因为多年以前心理学家确实对超感官知觉怀有极大的兴趣，直到累积了大量的负面证据之后，这种兴趣才消退了。正如历史所展示的那样，研究课题通常不是由某个权威政府机构宣布停止的，它们只是在生存竞争的环境中被自然淘汰出局了而已。

在心理学领域里，超感官知觉从来没有被认为是一个不能研究的主题，这一事实是清楚并且公开的（Farha, 2007; Hines, 2003; Kelly, 2005; Marks, 2001; Milton & Wiseman, 1999; Park, 2008; Wiseman, 2011）。有许多研究超感官知觉的论文发表在专业的心理学刊物上。最近在2011年，APA核心杂志出版了一篇关于超验心理学效应的文章（Bem, 2011）。唉，不出所料，报告指出这些效应并不可靠

（Rouder & Morey, 2011; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011）。

那些在媒体上频频曝光的超自然心理学家总喜欢让人们觉得这一领域是崭新的，有惊人的新发现即将出现。其实，事实却没那么激动人心。对超感知的研究和当代心理学自身的历史一样久远，它并不是什么全新的研究领域。在心理学文献中，它也曾经像许多现在被认为是可行的主题一样被认真地研究过。然而，在正式心理学刊物上所发表的有关这一领域的许多研究结果都无法证明超感知的存在。在20世纪历经90多年的研究之后，我们仍然无法在控制实验条件下重复验证任何超感知现象。尽管过去几十年来进行了大量有关超感知的研究，却从来没有一个研究能达到这一简单而基本的科学标准。这一点甚至连超自然心理学家及其信徒都承认。简而言之，尚未出现需要科学解释的未经证实的现象。仅仅这个原因，就使得心理学对这一话题失去了兴趣。

颇具讽刺意味的是，心理学家在评估超常能力方面扮演着关键角色。他们的重要性可能仅次于那些拆穿无数超能力演示骗局的专业魔术师（Randi, 2011）。而且，很多论述和质疑超常能力的重要书籍都出自心理学家之手。

讽刺意味显而易见。心理学作为一门最可能精确评估超感知言论的学科，在公众的心目中却与伪科学关系最近。这种“被连累”的现象让心理学深受其害。正如下面还要再详细讨论的，心理学常常会陷入这样一种“里外不是人”的境地，这只是其中一例。那种认为在心理学里没有什么规则、这个领域的知识缺乏科学评判标准的信念，导致人们将心理学与超感知这样的伪科学联系在一起。然而，如果心理学家成功地让公众认识到这些伪科学的真面目，心理学与伪科学的联系又会被视为“心理学不是一门科学”的铁证！

自助类读物

书店里常见的第三类心理学读物就是所谓的自助类读物。当然，这类读物也有许多不同种类（Lilienfeld, Lynn, & Lohr, 2003; Meyers, 2008）。有一些书是励志类的，目的是为了提升人们的自我价值感和自信心；另一些书则是新瓶装旧酒，将一些关于人类行为的老生常谈重新包装了一下；只有少数（简直是凤毛麟角）书籍是由负责任的心理学家为公众撰写的。还有许多书，虽出自心理学专业人士之手，但算不上“负责任”的作品。为了标榜其“独特性”，声称自己发

明了一些新“疗法”，不但可以矫治某些特殊行为问题，而且还能满足老百姓的一般需要（赚钱、减肥和拥有更好的性生活是其“三大”主题），这类书籍常能大卖。这些所谓的新疗法很少基于控制实验的研究，如果作者是个临床医生的话，他们通常只是依靠他们的个人经验或者少数的几个病例，就提出了自己的“发现”。这在所谓“替代性医疗”中十分常见。

有许多有效性经过了严格的心理学检验程序验证的认知和行为疗法，却很少出现在书店的货架上。利连恩菲尔德（2012）估计每年有3500种自助读物出版，其中只有5%的书中具有科学效度。

这一情况在数字化媒体和互联网上更为糟糕，电台和电视台几乎没有任何正规的心理学报道。相反，他们总是邀请一些江湖术士和爱出风头的媒体名人，而这些人与真正的心理学毫无瓜葛。媒体之所以会这样做是因为正规的心理疗法从来都不会声称自己能立竿见影、药到病除，甚至不会担保治疗一定成功，或者夸大其治疗的范围（如“你不仅会把烟戒掉，而且生活的方方面面都会得到改善！”）。

同样，现在互联网上也出现了类似的情形由于缺乏严格的同行评审，人们在网上看到的治疗方法通常都是骗局。这里有个例子。保罗·奥菲特（Paul Offit）2008年出版了一本名为《孤独症的错误预言者》（Autism’s False Prophets）的书，他详细描述了很多治疗孤独症的方法，这些方法虽然已经被真正的科学研究证明是虚假的，但在急于救治孩子的家长中却十分流行，其中一个是已经在第6章中讨论过的辅助沟通。奥菲特描述了很多其他伪科学的方法，这些方法错误地抬高了家长们的希望，让他们花费大量金钱和精力寻求一个虚假的“治疗”。在2012年1月5日，我认出了在奥菲特书中讨论过的一个虚假的治疗孤独症的化学“疗法”（在此我不提及其名字，以免替它做了广告），并在谷歌搜索中输入它和“孤独症”。在我搜素结果的前10个链接中，3个链接都指向鼓吹这一疗法的网址。网络搜索无法保证科学准确性，因为网站没有同行评议。随机的搜索用户对与问题主题相关的科学文献一无所知，网络搜索也无法为其提供消费者保障。

这类在美国图书市场中占相当比重的自助类读物，极大地影响了公众对于心理学的印象。首先，像弗洛伊德问题那样，这些书使公众搞不清心理学研究关注的焦点在哪里。举个例子来说，虽然有相当数量的心理学家在为肥胖、人际关系和性问题提供治疗，并在不断地进行研究，但这个数量比起自助类读物中所说的要少得多。这种误解也使得公众以为大多数的心理学家都致力于异常行为的研究和治疗。事实上，大部分的心理学研究的是人类的正常行为。

除了引起对研究内容的误解之外，自助类读物还让人们对心理学的研究方法和目的产生错误印象。正如第4章中所讲的那样，心理科学并不认为几个个案研究、见证叙述和个人经验就能构成支持某种疗法有效性的充分的实证证据——而这些却恰恰是大多数自助类“疗法”安身立命的根本。自助类读物因此误导了社会大众，使他们认为大多数的心理学理论就是基于此类证据得出的。在第8章中我们已阐明，证实一个理论需要许多不同类型的证据来支持，个案研究所提供的数据的说服力在其中是最弱的。将此类证据视为证实某一理论或疗法的确凿证据，无疑犯的是根本性的错误。

菜谱式知识

最后，自助类读物使公众误解了心理学的目标和多数心理学研究所追寻的知识。这种读物带给人们一种强烈的暗示，那就是认为心理学研究者所追求的是那种“菜谱式”的知识。菜谱式知识是指那些只告诉你如何去使用某物，但对其基本的运作原理一概不谈的知识。例如，大多数人知道如何使用电脑，但他们对电脑如何运作知之甚少。这就是电脑的菜谱式知识。在我们的社会里，许多有关科技产品的知识都是菜谱式知识。

当然，这也不完全是一件坏事。事实上，多数技术产品的设计初衷，就是为了让那些对其背后的运作原理一无所知的用户也能使用。事实上，菜谱式知识这一概念提供了一种方法，可以概括基础研究和应用研究之间的区别。基础研究工作者寻找自然界的基本原理，而不去考虑这些原理能否转化为菜谱式知识。应用研究工作者则致力于将基本原理转化成一个个只需菜谱式知识就能使用的产品。

多数自助类读物只提供关于人类行为的菜谱式知识，它通常能够简化为这样的形式：“你只要做X, 你就会变得更加Y了”，或者“做Z, 某A就会表现出更多的B”。当然，如果这个药方是正确的（这一假设往往并不全然成立），这么做也不为过。许多正规的心理治疗都提供了大量菜谱式知识。然而，当人们错误地认为所有心理学研究的终极目标就是提供菜谱式知识时，问题就产生了。尽管许多心理学研究者确实致力于将基本的行为理论转化为实用的心理疗法、保健行动方案或有效的工业组织模式，但心理学主要还是一门发现行为的普遍事实和理论的基础学科。这就是心理学研究为何会让外人觉得很怪的另一个原因：基础理论的研究与应用研究之间存在巨大的差异。

如果一个人走进分子生物学实验室，并询问一位研究者：“我们

在头痛时是应当服用两片还是三片阿司匹林？”我们会觉得这个人很傻。原因并不在于分子生物学与缓解疼痛没有任何关系，事实上对止痛药的研究可能会运用到这一领域的知识。我们之所以说这个问题问得傻，是因为分子生物学家并不是那种在开药方层面上工作、回答你是要吃两片还是三片阿司匹林的人。研究者所关注的是有关生物成分在分子水平上的基本数据。这些数据可能会为许多不同领域提供菜谱式知识，但发现基本数据和将这些数据转化为菜谱式知识的人不大可能是同一个人，而转化为菜谱式知识的方法，也会与最初发现事实的方法有所不同。

由于自助读物让公众错误地相信，多数心理学家都致力于开发菜谱式知识，这使得许多心理学家所做的基础研究显得颇为奇怪。海奇特曾让被试在一间黑暗的房间里注视着一个小红灯（第7章），这到底与我们的现实世界有什么关系？是的，从表面上看来，确实没有一点儿关系。海奇特是对人们的视觉系统如何适应黑暗的基本原理感兴趣，这些基本原理最终会转化成能够用来应对具体问题的菜谱式知识，例如夜盲症是由维生素缺乏导致的。然而，这一转化并不是由海奇特本人来完成的，而且它在几年之后才到来。

因此，自助读物给公众对心理学的感知带来两种不良的副作用。第一，这些读物中涉及的问题并不能代表当代心理学关注的焦点，相反，它们通常反映的只是消费者想看的内容。心理学学生往往无法充分地意识到，图书出版是一种商业行为，市场的力量决定了什么样的读物可以摆在书店的书架上。然而，科学的关注点并不是由此决定的。在所有的学科尤其是心理学中，被科学家们认为有用的想法和那些被包装后能够热销的想法之间存在一道鸿沟。例如，在心理学上有一些关于“积极思维的力量”的合理研究（Sharot, 2011），但它和《奥普拉脱口秀》上听到的自助处方的效果几乎没有什么重合之处。相反，心理学研究的文献上充斥着各种注意事项、关心聚合证据并寻求研究方法上的连通性——简而言之，在这本书中阐述的所有真正研究所关心的事情。

再说说减肥药物领域。科学家们慢慢地掌握了一些温和的药物可以帮助人们控制体重的累积性证据（Brody, 2008），但这些绝非突破性的治疗手段。很明显，肥胖问题很复杂，并且受制于我们警告过的多重因果（Bartoshuk, 2009）。这个问题显然不会被单一的神奇疗法所解决。例如，很多科学家强调，食物环境本身的复杂性（广告、食物分量、儿童市场）在一定程度上诱发了民族肥胖问题（Brownell, 2011）。

作为对比，来看看基础医学博客的作者、退休医生哈莉特·霍尔

（Harriet Hall, 2008）的一份报告。她提到了一种减肥产品“作出了异乎寻常的许诺：随便吃，还能瘦。但它有着绝佳的广告语：‘如果不是真的，我们绝对不会印出来！’这太好笑了。任何人无论说什么都能印出来，除非他们被逮到。这些减肥广告都在信口开河，联邦贸易委员会不能将他们都抓起来。”（p.47）。霍尔要说的重点是，真正的科学与媒体（从电视、印刷品到网络）想让公众知晓的完全没有联系。媒体想要快速回答公众感兴趣的问题，而科学对科学可解问题产生答案的过程比较缓慢，并且公众感兴趣的问题可能是无解的。

心理学和其他学科

当然，心理学并没有垄断对于行为的研究。许多其他的相关学科采用不同的技术和理论视角，也对我们关于行为的知识有所贡献。许多涉及行为的问题都要求多学科的取向。然而，大多数心理学家必须要接受的一个非常残酷的事实就是，当这种多学科问题的研究成果发表时，心理学家的贡献往往会被其他学科所掩盖。

关于心理学家的贡献被忽略、抹杀或者被部分归为其他学科的例子不胜枚举。例如，第一个有关电视暴力对儿童行为影响的研究是由美国公共卫生局主持的，研究结果发现二者之间存在因果关系，因此，之后由美国医学会通过一项决议，重新确认该项研究的成果并向公众推广。这本来也是顺理成章的事，的确没什么错，但这一举措在无意间造成了一个后果，就是媒体不断地将电视暴力的研究成果与美国医学会联系在一起，给公众造成了这样一个印象，即确立这一发现的研究是由医学专业人士主持的。事实上，绝大多数有关电视暴力对儿童行为影响的研究都是由心理学家完成的。

另一个导致心理学家的工作经常被划入其他学科的原因是，这些年来，“心理学家”一词的含义已经含混模糊了。许多心理学研究者在标识自己时，往往把自己的研究专长加在“心理学家”之前，例如自称生理心理学家、认知心理学家、工业心理学家、进化心理学家和神经心理学家。还有一些称谓甚至摒弃了“心理学家”一词，例如神经科学家、认知科学家、社会生物学家、人工智能专家和行为学家，等等。所有的这些举动，再加上媒体认为“心理学不是一门科学”的偏见，都导致了心理学家的成就被误划入其他学科：生理心理学家的成果被划归生物学，认知心理学家的成果被归为计算机科学，工业心理学家的成果被归入工程学和商学，等等。即使当代最杰出的心理学研究者之一丹尼尔·卡尼曼获得了2002年的诺贝尔经济学奖，心理学也没有分享到任何好处！当然，诺贝尔奖中没有为心理学单独设立奖项（Benjamin, 2004; MacCoun, 2002）。

事实上，忽略心理学的倾向是十分荒谬的。2008年4月17日，《纽约书评》在86页刊登了以下一则更正：“在休·哈尔珀林（Sue Halperin）对幸福感（纽约书评，4月3日）的评论中，经济学家丹尼

尔·卡尼曼做出的先驱性研究应该是快乐心理学（hedonic psychology），而不是享乐心理学（hedonistic psychology）。”最初我们可能认为杂志很精准——他们将错误的享乐心理学更正为快乐心理学。然而，编辑没有注意到，在印这份更正前，他们又犯了另一个错误——丹尼尔·卡尼曼不是经济学家，而是认知心理学家！

埃莫里大学耶克斯灵长类动物研究中心的负责人、心理学家弗雷德里克·金（Frederick King）曾讲到，某天他花了大段时间解释动物模型对人类神经障碍研究的重要性，在聆听完这位在癫痫症的神经和行为研究方面成就斐然的学者长时间讲解之后，有位记者问道：“你不过是个心理学家，怎么会知道这么多关于癫痫症的事儿呢？”

最后，想想2007年在前白宫助手刘易斯·（小摩托）·利比[Lewis（Scooter）Libby]身上发生了什么。来自一名著名研究型心理学家的专家证词未被采纳，因为法官认定，众所周知，记忆是不可靠的，陪审团能够依据他们的常识去判断记忆是如何工作的。事实上，研究表明大约30%的群众相信人类记忆“像磁带录音机一样工作”（Lilienfeld, 2012）。与法官认为的相反，30%的陪审团非常需要听从专家的！

我们是自己最坏的敌人

我们不是只会把心理学的形象问题怪罪在其他人头上，心理学家自己在这方面也“功不可没”。由于试图把真正的心理学介绍给公众的正规心理学家往往得不到什么好的回报，大多数研究型心理学家和公众缺乏交流。

尽管如此，美国心理学会（APA）和美国心理协会（APS）正致力于促进与公众的沟通（West, 2007）。美国心理协会为此还新创办了一本期刊，名为《公众感兴趣的心理科学》（Psychological Science in the Public Interest）。美国心理协会也为此开了一个叫作“我们只是人类而已”（www.psychologicalscience.org/onlyhuman）的博客。心理学需要在这一方面再加把劲。不然的话，对于公众对我们学科的误解，我们就只能怪罪自己了。

美国心理学会前主席罗纳德·福克斯（Ronald Fox）在最近的致辞中谈到了心理学在沟通和传播方面的问题，以及我们自身是如何带来这些问题的：

> 一些经常在大众传媒上露脸的从业者，他们的做法是不专业的、不道德的，并且使他的同行蒙羞……我们的学科对于那些不负责任的、令人发指的公开欺骗缺乏有效的对策……当今世界里，公众成天接触的观点和意见都来自于一些骗子（在最近的一个电视脱口秀节目里，一个心理学家声称他已经帮助许多病人回忆起前世所受的精神创伤），而不是理性的心理学从业者。（Fox, 1996, pp. 779-780）

最后，心理学的某些分支中存在着一些反科学的态度和现象。例如，在一些心理治疗的圈子里，有人一贯拒绝对自己所采用的疗法进行科学评估。专栏作家和心理治疗师查尔斯·柯瓦斯阿默（Charles Krauthammer, 1985）写了一篇文章，论述了这种态度对心理治疗的声誉造成的严重损害。第一，由于拒绝去莠存良，造成各种疗法泛滥成灾。这种泛滥不仅使消费者的权益受到损害，而且还加深了这一领域的误区。柯瓦斯阿默慧眼如炬，看出不遵循证伪原则的做法已经阻碍了科学进步。

柯瓦斯阿默最后又指出了心理治疗这个圈子的一个内在矛盾：一方面，他们认为心理治疗“更像一门艺术而非科学”，因此反对以科学的方法进行评估；另一方面，他们仍然非常关注所谓的“800磅大猩猩”[1]，即政府补贴和个人健康保险。柯瓦斯阿默揭示了心理治疗圈子中这两种态度的矛盾性：如果心理治疗真的是一门艺术的话，他们应当由国家人文基金提供资助，而不是医疗保险。

本书早期版本的一些读者指出，我并没有特别强调心理学家内部的不专业行为和反科学态度在很大程度上导致了这个学科的公众形象问题，因此指责我“轻易地放过了心理学家”。为均衡起见，我借鉴了许多罗宾·道斯（1994）和斯各特·利连恩菲尔德（2012）的工作。谁若是对“心理学家自身就是造成此困境的重要原因”心存疑问，那就去读读这两位学者的书。道斯毫不犹豫地揭下心理学的遮羞布，并主张在专门研究人类问题的心理学里，采取科学态度对于整个社会有很大的实用价值（虽然其潜力仍大有可挖）。例如，道斯认为：“确实有一门真正的心理科学，这门科学是在无数人多年以来的工作基础上发展而来，但是，这门科学目前正因为一些从业者的行为而逐渐被忽视、贬低和遭到反对——这些从业者只是在口头上承认这门科学的存在而已。”（1994, p.vii）与之类似，利连恩菲尔德（2012）认为“心理学家应该避免把心理学形象受损的所有责任都推到公众的普遍误解身上的肤浅诱惑。至少一些心理学的污名罪有应得，这个领域林子太大，尤其是从属于心理治疗的，依然深陷不科学的做法中”（pp.122—123）。

道斯和利连恩菲尔德所反对的是心理学领域基于心理学的科学地位颁发资格认证，然后又用资格认证来保护心理学从业者的不科学行为。例如，一个受过良好训练的心理学家应当知道，我们有把握对总体的行为作出预测，但是在预测某个特定个人的行为时，就存在很大的不确定性（见第10章和第11章），因此，即便是最有能力的心理学家，也不应该在没有强调这点的情况下去作任何个人预测。正如道斯（1994）所言：

> 一个声称有百分之百把握预测某一个体未来行为（如暴力行为）的专家，注定不是一个称职的专家。因为有研究证实，不论是一个心理健康专家，或者其他什么人，都不可能以这样的把握保证自己预测的准确性（专业人士经常声称，尽管他们就个人来讲是接受不确定性的，但他们的专业角色“要求”他们做出这样自信的判断。不，他们不是被“要求”这样做，是他们“自愿”这样做的，p.vii。

简而言之，美国心理学会曾经助长了心理治疗领域的这股不正之

风。这股风气让人觉得，心理学家能够通过训练获得一种“直觉洞察力”，从而能洞悉个体的行为。然而，研究证据并不支持这一观点。当有人提出质疑，认为执照制度只是一种行业限制时（该组织就把它的科学资历作为武器），一位美国心理学会主席这样回应社会人士对心理学的攻击：“我们是以科学为基础的，这就是我们有别于社会工作者、咨询师和吉普赛卜卦者的地方。”（Dawes, 1994, p. 21）但是，用来维护其科学地位的这个理由却正好揭示了有执照的心理学家具有独特的“临床洞察力”的观点是完全错误的。美国心理学会这种两面派手法催生了道斯的这本书，也在一定程度上导致了20世纪80年代美国心理协会（APS）的成立。这一协会的成员是由那些厌倦了美国心理学会“只关注蓝十字津贴而忽视科学”的做法的心理学家所组成。

斯科特·利连恩费德——一位因其事业早期对临床心理学所做贡献而获得大卫·沙科夫（David Shakow）奖的学者，曾在颁奖典礼上不断重申上述观点，并警告说：“在临床心理学这一领域，我们似乎对处理伪科学这一问题完全没有兴趣，这是一个非常令人吃惊的现象，因为这个问题的火苗已经烧到我们的后院了。”（Lilienfield, 1998, p. 3）他还列出了20世纪90年代在临床心理学领域泛滥成灾的几种伪科学，其中包括：用于治疗创伤的那些未经检验的怪异疗法；已经被证实是无效的、针对孤独症的一些疗法，例如辅助沟通疗法（见第6章）；继续使用的一些未被充分验证的心理评估工具（例如各种投射测验）；潜意识自助录音带；使用高度暗示性的治疗技术诱发儿时受虐的记忆。

利连恩费德援引著名临床研究者保罗·米尔（Paul Meehl）的话：“如果我们不对这一行业进行清理整顿、为我们的学生提供科学思维典范的话，外行就会替我们做。”（p.728）米尔在此指出了我们在第11章中讨论过的一种倾向：临床治疗师总是想让别人相信，他们拥有一些关于人类的“特殊”知识，这些知识超越了公众可获悉的、作为可重复验证的科学知识的一般行为趋势。米尔（1993）认为，临床心理学家必须更关注那些实证的可公开验证的知识，并警告说，“如果认为自己拿到了博士头衔，就自以为能够在取样、感知、记录、保持、提取和推断这些人类心理受限的方面不犯错误，这是非常荒唐和自大的”（p. 728）。

然而，心理学领域依然在遭受不端行为的践踏。例如，“紧急事件应激晤谈”在许多场合被作为标准化的程序，用于治疗那些经历了爆炸、枪击、战争、恐怖主义和地震的患者（Groopman, 2004; McNally, Bryant, & Ehlers, 2003）。晤谈程序包括让患者“谈论事件并公开表达他们的情绪，尤其是当着也经历了同样事件的公司同事的

面”（McNally et al., 2003, p. 56），其目的是为了减少创伤后应激障碍（PTSD）的发生。大多数经过晤谈的病人都报告说这种体验是有帮助的。当然，看过此书的人都不会认为其依据具有说服力（想想第4章中关于“安慰剂”效应的讨论）。显然，需要有一个控制组（不给予危机事件压力舒解）。事实上，“许多创伤幸存者都在没有专业帮助的情况下从最初的创伤后反应中恢复了过来”（McNally et al., 2003, p. 45），因此需要证明，重大事件应激叙事的使用确实带来了更高的恢复率。虽然真正的控制实验所揭示的结果并非如此（Groopman, 2004; McNally et al, 2003），但这一疗法还在被继续使用。

艾莫瑞、奥托和奥多诺胡（Emery, Otto, & O'Donohue, 2005）在搜集大量证据后所做的综述中指出，与儿童监护权相关的临床心理学中充斥着伪科学（Novotney, 2008）。例如，研究者描述了一些临床心理学家在儿童监护权官司中惯用的用以评估儿童最大利益的工具。在回顾了此类工具——例如，传说能够测量关系知觉和父母觉知能力的量表——之后，艾莫瑞等人（2005）得出结论：没有一个工具被证明是可靠而有效的。他们写道：“没有一个关于这些测量方法有效性的研究发表在具有同行评议机制的刊物上，而这是科学的一项重要标准。”（p. 8）同时他们总结道：“我们对于这类测量最保守的评估也是尖刻的，即这些测量的结构界定不清，并且表现得如此糟糕，在儿童监护权评估中的运用未经任何科学的检验。”（p. 7）

艾莫瑞等人（2005）指出，不仅评估儿童监护权的工具存在缺陷，而且临床心理学家使用的概念也有问题。艾莫瑞等人举了一个所谓“双亲疏远综合征”的例子。这个概念完全基于单独个案的“临床经验”，并且缺乏科学研究结论所需要的聚合效度，但它在监护评估中却被临床心理学家当作真正的科学概念一般随心所欲地使用。类似的是在对性侵犯的评估中也有这样著名的测量方法。尽管他们缺乏预测效度——测量方法没有能力区分性地预测再次侵犯的可能性，但临床心理学家仍然在使用它们（Ewing, 2006）。类似地，大部分被临床医生用来预测心理变态的人的未来暴力行为的手段实际上并没有声称的那样准确（Skeem, Polaschek, Patrick, & Lilienfeld, 2011; Yang, Wong, & Coid, 2010）。

不过事情似乎有了一些转机。2002年一本新的杂志诞生了，它就是《心理健康实践的科学述评》（The Scientific Review of Mental Health Practice）（Lilienfeld, 2002, 2007; Lilienfeld, et al., 2008）。这本杂志致力于区分科学的治疗方法与那些伪科学的治疗方法，它已经得到科学心理健康实践委员会的认可。更令人振奋的是，

至少有一些心理学组织已经痛下决心来整顿临床实践，并准备消除在实践过程中那种根深蒂固的“什么都能往里装”的态度。利连恩费德和洛哈（Lilienfield & Lohr, 2000）报告了亚利桑那州心理学资格审查委员会吊销一位心理学家执照的事件。这个心理学家试图以一种伪科学的治疗方法来治疗恐惧症，这种方法是按照预定的顺序拍打患者身体的各个部位。不用说，这种方法没有实证效度。亚利桑那州委员会命令该治疗师停止使用这种方法，并且给他“留职察看”的处罚——一个心理学组织对使用伪科学方法的成员进行查处，这样的例子在心理学界还是非常罕见的。

2009年，心理科学协会制作一篇重要的报告阐述了临床心理学的状况，得出的结论是“临床心理学在其历史中的某一点上类似于医学，那时职业医生用前科学的方式做手术。在19世纪早期医学改革之前，医生通常和很多现代的临床心理学家拥有一样的观念，比如说重视个人经验而不是科学研究……大量的证据显示，很多临床心理学博士的培养项目，尤其是心理学博士和以营利为目的的项目，都存在没有对毕业设置高标准的门槛、师生比过高、培训中忽略科学性、培养出的学生不能应用或者产出科学知识的问题”（McFall, & Shoham, p.67）。这篇文章得到了很多关注，但是大众媒体的一些讨论对这一问题的混淆大过澄清。另外，在《新闻周刊》刊登的一篇失之偏颇的报告不幸被命名为“忽视证据：为什么心理学要拒绝科学？”（Begley, 2009）。文章错误地认为，所有的心理学拒绝科学，而不是临床心理学中有问题的分支。这个有迷惑性的标题带有苦涩的讽刺意味，它假定APS报告的逻辑是，所有坚持科学方法心理学的其他分支，在悲苦地向唯一不坚持科学方法一个分支（临床心理学）喊话。

简而言之，心理学具有像吉柯（Jekyll）和海德（Hyde）[2] 那样的双重人格，极端缜密的科学与伪科学及反科学的态度并存。这个学科的双重人格特性在近20年关于“恢复记忆—虚假记忆”的争论中表现得淋漓尽致（Brainerd & Reyna, 2005; Gardner, 2006; Lilienfeld, 2007; Loftus & Guyer, 2002; McHugh, 2008）。许多个案报告说，有患者声称回忆起几十年前当他们还是小孩的时候遭受虐待的经历，而这些记忆曾经一度被遗忘。大部分这类记忆出现在治疗干预的情境中，显然说明这些记忆中的一部分是由治疗本身所引发的（Gardner, 2006; Lilienfeld, 2007; Loftus & Guyer, 2002; Lynn, Loftus, Lilienfeld, Lock, 2003）。有些人坚持认为这类记忆绝对不可信，另外一些人则坚称它是可信的。在这个爆炸性的社会话题所营造出的极具情绪化的氛围下，心理学家们提供了一些较为理性、平衡的意见，更为重要的是还提供了部分关于恢复性记忆或虚假记忆的客观的实证证据（Brainerd Reyna, 2005; McNally & Geraerts, 2009; Moore &

Zoellner, 2007）。

从这里，我们能充分地看出心理学这种双重人格的特性。由治疗干预所引发的、与事实真相相反的虚假记忆中，有一部分是由某些不称职的、对科学无知的治疗师造成的，而这些治疗师都是心理学专业人士。另一方面，尽管目前对这场争论所做的结论还不够充分和确定，但这一点仍应归功于那些对相关现象实证地开展研究的心理学家的不懈努力。最后，我必须说明不是只有心理学才有这样的困扰。实际上，医学一直在踢打叫喊着朝着完全基于证据的方向蹒跚前行，并且现在仍然还在路上（Gawande, 2010; Kenney, 2008）。

我喜欢引用丹杰费尔德的口头禅来作为本节的题目，希望这样做能够帮我洗清“为心理学家脱罪”的恶名。心理学家道格拉斯·穆克（Douglas Mook）在他的一本关于研究方法的书中曾提到过我借用丹杰费尔德的笑话，并且评论道：“确实，通常心理学得不到应有的尊敬，但有时，它又受到了不应得的尊敬，或者因为错误的原因而受到尊敬。”（Mook, 2001, P.473）我完全同意这一感受。穆克是对的，心理学的学生应当知道这个学科所面临的窘境。就像本书中所表述的那样，作为一门研究人类行为的科学，心理学通常没有得到太多的尊敬。但是，心理学呈现给公众的印象却是很多临床治疗师宣称自己具有“独特的”洞察人心的能力——但这种洞察力在研究证据方面是站不住脚的，这一形象又使心理学获得了过多的尊敬。心理学的严谨性就在于，采取科学的方法来验证有关人类行为的各种主张，不幸的是，这一学科常由那些不尊重心理学这一严谨性的分支呈现给公众。

注释

[1] 指具有压倒一切的影响力的事物。——译者注

[2] 英国作家史蒂文森的小说《化身博士》中的人物，分别代表善与恶。——译者注

每个人不都是心理学家吗

我们每个人都有一套关于人类行为的理论。很难想象，如果没有这些理论，我们该怎样活下去。从这一意义上讲，我们人人都是心理学家。尽管如此，区分这种个体心理学和由心理科学所生成的知识体系仍然十分重要。我们将看到，这种区分之所以重要，是因为二者的区别在许多大众读物里经常被故意混淆了。我们的个人心理学知识与那些对行为进行科学研究所获得的知识相比，有哪些方面的区别呢？我们已经有所讨论。我们的个人心理学知识多数是“菜谱式知识”。我们做某件事，是因为我们认为它会导致其他人做出某些相应的行为，或是因为我们相信这些事能帮助我们实现某些目标。这些都是所谓的菜谱式知识。但是，个人心理学和科学心理学（也包括一些菜谱式知识）的区别并不在于有没有菜谱式知识。最主要的区别在于，科学心理学总是力图通过实证方法检验菜谱式知识的有效性。

科学评估具有系统性和可控性，这些特性是个人评估程序所不可能具备的。事实上，心理学对于决策选择的研究表明，当行为发生的情境与原有的信念相悖时，人们就很难觉察到相关关系（见Baron, 2008;Stanovich, 2009）。我们只看到我们想看到的东西。心理学家已经找到出现这种现象的许多原因，但是它们并非我们这里关注的重点。即使我们想在个人的基础上评估个体的菜谱式知识，那些妨碍我们对行为现象进行充分观察的先入为主的偏见也会使我们的评估工作变得异常困难。引入科学方法的目的正是要避免个别观察者的偏见。这里的意思很简单，由科学心理学产生的菜谱式知识可能会更精确，因为和个体的菜谱式知识相比，它们经过了更加严格的检验程序。

就像本章前面所讨论的那样，个体心理学和科学心理学之间的差别不仅限于对菜谱式知识的验证。科学想从自然界获得的远不止菜谱式知识。科学家们想要寻求那些能够解释药方运作机制的更为普遍的基本原理。许多人的个体心理学和科学心理学一样，也想探究更为基本的心理学规律和理论，然而这些个人理论和科学理论存在着重大的分歧。我们曾经提到过，这些个人化理论是无法证伪的。许多人的个人心理学理论缺乏缜密的建构，只是一些适用于个别情形的陈词滥调的简单堆砌，有时这些话还会自相矛盾。它们向人们保证，存在一个确定的解释，而那些与之完全对立、会彻底动摇人们信念的事件都是

不可能发生的。尽管这些理论极具慰藉功能，但正如第2章中所讨论的，除了慰藉之外，以这种方式提出的理论再无其他功能。这些理论都以“事后诸葛亮”的方式解释一切，对未来没有任何的预测。没有预测，也就没有给我们提供任何信息。心理学科的理论必须符合可证伪的标准，这就是心理科学与许多外行人的个人心理学的不同之处。心理学理论是能够被证伪的，因此，心理学理论蕴涵了这样一种确保其发展和进步的机制，而这是个人心理学所不具备的。

抵制科学心理学的根本原因

基于我们之前讨论过的那些理由，千万不要把个人心理学理论和科学心理学的知识混为一谈。这种混淆有时是蓄意制造出来的，目的是要诋毁心理学在公众心目中的形象。如果“人人都是心理学家”是指每一个人都有自己的心理学理论的话，那么这句话没有错。但是它常常被隐晦地暗示心理学不是一门科学。

在第1章中已讨论过，为什么科学心理学的想法会对某些人造成威胁。一门日趋成熟的行为科学，势必会改变各类提供心理信息数据来源的个体、群体和组织。很自然，对那些长期从事人类心理和行为评论的人来说，他们肯定会抵制任何威胁其权威地位的变革。在本书的第1章中曾提到过，科学的进步会不断地剥夺那些原有对自然界作出解读的权威团体的地位。行星的运行、物质的本质、疾病的原因过去曾经是神学家、哲学家和通才作家把持的领域，而如今，天文学、物理学、医学、遗传学和其他学科逐渐夺取了这些主题，并将它们放置在不同的科学专门领域内。

举例来说，许多宗教都已经逐渐不再声称他们对宇宙结构具有专门的知识。除了一些局部性的争议——如特创论——科学与宗教之间的大型战争已经成为历史。科学家们探究自然世界的结构，而许多宗教则对运用这些发现时可能带来的影响作出评论，但宗教已经不再与科学争夺对于这些发现的解释权了，对有关自然界的主张的裁定权，无疑已经掌握在科学家手中。

作家纳塔莉·安吉尔（Natalie Angier, 2007）提醒我们，很多年前，当闪电击中教堂的木质尖顶并将其烧毁，神职人员和老百姓都会加入一场激烈的辩论：这是不是“上帝的惩罚”的征兆。但提醒我们“在18世纪，本杰明·富兰克林发现闪电只是电而不是神迹。他建议将导电棒安装在所有的尖顶和屋顶上，之后关于雷霆之箭的辩论也就烟消云散了”（p.26）。

接下来的问题就是信念评估标准的变革。不会再有新闻报纸刊登有关土星带构成的立场鲜明的社论文章。为什么呢？并没有审查机构阻止这类社论的发表。很明显，写这类社论是徒劳的。因为社会大众

知道，对这一方面的知识有发言权的是科学家，而不是评论员。仅在一百年前，报纸和那些布道坛上的牧师还曾对动物世界的物种起源学说大肆攻击。现在，这类评论大部分都消失了。科学摧毁了让任何理性思考者轻信这些观点的客观条件。心理学还将在另外一个庞大的自然领域中摧毁这类条件。

有些人发现自己很难接受心理学发生的这类变革。他们顽固地坚持自己有权利对人类行为发表看法，即使这些看法与事实相去甚远。显然，“权利”用在这里并非是一个准确的措词，因为在一个自由社会里，每个人都有发表意见的权利，无论这些意见是否正确。最重要的是，要意识到许多人想要的不仅仅是发表有关人类行为见解的权利，他们真正想要的是，无论他们说什么，人们都应该相信其所说的话。当他们陈述一个关于人类心理学的观点时，他们希望周围的环境有利于人们接受他们的想法，这就是为什么认为心理学是“什么都能往里装”的说法会有大量拥护者的原因。所谓“什么都能往里装”，就是暗含心理学的主张是不能由实证方法来判别的，它只是一堆观点的集合。科学对于这种“什么都能往里装”的观点来说始终是一种威胁，因为它有一系列严格的标准和程序，用以确定哪些说法是可信的。科学不是“什么都能往里装”。正是这种去伪存真的能力推动了科学的进步。

简而言之，许多对于科学心理学的抵制都可以归因为“利益冲突”。前面几章中已经讨论过，许多伪科学已经发展成为数以百万美元计的产业，它们之所以能蓬勃发展，依靠的正是公众没有意识到关于行为的主张也可以用实证方法来检验这一事实（在美国，占星师的数量是天文学家的20倍；见Gilovich, 1991, p.2）。公众也没有意识到，支撑这类产业的许多主张（如星相预测、潜意识减肥、辅助沟通及通灵手术）都已经被证明是无效的。美国国会下设的一个委员会曾经估计过，人们每年在这些医疗骗术上大约花费100亿美金，这让花在正规的医疗研究上的经费相形见绌（Eisenberg et al., 1993; U. S. Congress, 1984）。

我们如何识别伪科学的主张？临床心理学家斯科特·利连恩费德（Lilienfeld, 2005, p. 40）给出了一些注意事项，也是对本书内容的一个概括。他认为伪科学的主张有以下一些特征：

- 喜欢采用特殊的假定，使得其主张免于被证伪；
- 强调主张是确证的，是不可辩驳的；
- 喜欢将提供证据的任务强加给怀疑者，而非拥护者；

- 过度依赖轶闻趣事和各类见证叙述来证实其主张；
- 逃避同行评审；
- 并非建立在已有的科学知识之上（缺乏学科关联性）。

真正的科学家会苦心孤诣地强调这些标准，而不是回避它们。例如，三个致力于将情绪智力（EI）引入心理学的科学家就曾担心，媒体、临床医生甚至其他一些研究者会以非科学的方法运用这一概念。他们写了一篇文章，专门引导其他人使用上面列出并在本书中讨论过的科学标准："在我们看来，媒体普及情绪智力，总是频繁使用要么不完整要么过分宽泛的定义、不切实际宣称和对概念和研究更广泛的误解。我们强烈要求研究者和从业者参考关于情绪、智力和情绪智力的科学文献以规范他们的思想。简而言之，研究者需要引用科研文献，而不是新闻工作者对概念的理解，这两者服务于完全不同的目的"（Mayer, Salovey, & Caruso, 2008, pp.513—514）。

相比之下，许多伪科学术士和治疗骗术靠的就是心理学领域这种"什么都能往里装"的氛围。这是一种非常容易让公众变得盲信和盲从的环境，因为，如果"什么都能往里装"，公众的消费者权益就得不到保障了。正如律师彼特·哈勃（Peter Huber）所言："在科学的边缘和科学之外……形形色色的顺势疗法药物、水晶和金字塔神奇疗效的信徒……必须借助对正统科学的诋毁来为他们的异端邪说提供立足之地"（Huber, 1990, p.97）。这些兜售伪科学的人从骨子里想去掩盖这样一个事实，那就是有一套科学机制可以用于检验行为理论。

迈克尔·吉瑟林（Michael Ghiselin, 1989）警告说："道理很简单，人们都试图推销特定的观点，而真正能评估观点好坏的人，不是那些在市场上推销这些观点的人。"（p. 139）在行为理论和治疗这一领域，心理学家就是那些"知道如何来评估产品"的人。这就是为什么伪心理学产业一直极力反对科学心理学在评价有关行为的主张方面的权威性。然而，伪科学的散播者通常不与心理学家正面交锋，他们绕过心理学，带着其主张直奔媒体而去。大众传媒为那些想要绕过科学心理学的狂徒、骗子和伪科学提供了极大的便利。泛滥的电视脱口秀节目并不要求嘉宾出示科学研究的依据。这些嘉宾只要"足够有趣"，就可以在电视上露脸。在互联网上情况也好不到哪儿去。任何人都能建一个网站并且声称（贩卖）任何东西。退一步说，网站没有同行评审！

世俗智慧通常包含许多一厢情愿的想法：人们更愿意相信世界是他们所期望的样子，而非其真实的样子。为此，科学家们承担着费力

不讨好的任务，那就是去告诉公众：这个世界的本质并不是他们所想象的那样（“不，快餐对你的健康没有好处”）。媒体本来可以发挥有益的作用（告诉人们真相，而不是迎合他们的期望），然而，它们却把重心放在“娱乐”而不是提供信息上，从而使情况变得更糟。

科学确实是在把那些不符合最低检验标准的、自称是特殊知识的理论及疗法清除出局。法庭也在摒弃那些有关特殊知识的主张。在一桩著名的道伯特（Dauber）诉梅里尔·道（Merrell Dow）药品公司的公案中，最高法院决定何时才可以在法庭上呈现专家证词，也就是说，什么才可以让专家证词具有专家性！法庭为那些考虑专家证词的法官提供了四个鉴别指标：（1）观点所基于的理论基础是“可检验的”；（2）与某一方法关联的错误率可知；（3）观点所基于的技术或方法是否经过了同行评审；（4）技术或方法是否被相关的科学团体所广泛接受（Emery, Otto, & O'Donohue, 2005; Michaels, 2008）。这四个标准对应了本书的主旨：（1）可证伪性；（2）概率性预测；（3）服从同行评审的公共知识；（4）基于聚合性和共识的科学知识。法庭在排查特殊知识的主张、直觉和见证方面与科学相似。

本书曾经很简略地提及，在科学领域里的充分检验和不充分检验各是什么。内省、个人经验和见证叙述都被认为是关于人类行为主张的不充分检验。在科学心理学诞生之前，这些内容就一直被那些非心理学家的评论者视为支持其观点的宝贵证据，因此，此时会爆发冲突也就毫不奇怪了。

请不要以为我想把科学心理学刻画成一个充满敌意而让人扫兴的角色。恰恰相反，科学心理学的研究发现其实要比那些整天在媒体上反反复复、大呼小叫的伪科学有趣和精彩得多。因此，也不应该认为科学家是反对幻想和奇幻的，相反，当我们进入歌剧院或电影院时，总想看到幻想和奇幻，但这种情形却不太可能发生在我们去看病、买保险、到托儿所给孩子注册、坐飞机或者修理汽车的时候。这种情形也不太可能发生在我们进行心理治疗、将自己注意力缺失的孩子交给教育心理学家测试，以及把自己的朋友带到大学心理诊所进行自杀干预的时候。心理学在追求真相的过程中，必须像其他学科一样，把那些毫无依据的胡思乱想、“常识”、商业广告卖点、宗教意见、见证和妄想都清除出去。

让一门科学去告诉社会中的一部分人，他们的想法和意见是有用的，但不是在此地——这是一件相当困难的事情。心理学是科学中最后一个面临这种微妙局面的学科。这与心理学产生的时间有关。大多数学科成熟于精英控制社会结构的年代，那个时候普通人的意见没有

影响力。而心理学则产生于一个民主的传媒时代，忽视公众意见会危及自身。许多心理学家正在努力修复心理学与公众在沟通方面的糟糕记录。当越来越多的心理学家开始在与公众的沟通中发挥作用时，势必会加剧他们与那帮将个人心理学和科学心理学混为一谈的人的冲突。

尽管我们每个人都有一套直觉的物理学理论，但不是每个人都是物理学家。但是，如果我们不要求让个人物理学理论取代科学物理学，就为我们每个人获悉真正的科学物理学理论（因为科学是公共性的）铺平了道路。同样的道理，并非人人都是心理学家，但人人都能获得心理科学所发现的事实和理论，将之付诸实践，并丰富我们所有人对彼此的理解。

结束语

这本勾勒出“什么才是真正的心理学”的书至此已到了尾声。这一勾勒很粗略，但它对你理解心理学这门学科是如何运作的以及如何评估新的心理学主张应当有很大的帮助。我们勾勒的草图揭示了以下几点。

1. 心理学的进步是通过研究可解的实证问题而取得的。这种进步是不均衡的，因为心理学由许多不同的子领域构成，某些领域的问题要比其他领域具有更高的难度。

2. 心理学家提出可证伪的理论来解释他们的研究发现。

3. 理论中的概念都具有操作性定义，这些定义将随着证据的积累而逐渐演变。

4. 这些理论是通过系统实证的方法来检验的，用这种方法收集来的数据是公开的，也就是说，它允许其他科学家重复这些实验并提出批评。

5. 心理学家的数据和理论，只有在那些经过同行评审程序的科学刊物上发表之后，才算是进入了科学领域。

6. 实证主义之所以具有系统性，是因为它遵循控制和操纵的逻辑，这二者也是真实验的特性。

7. 心理学家采用许多不同的方法来获得他们的结论，这些方法的优缺点各有不同。

8.最终被揭示的行为规律，通常情况下都是一种概率关系。

9.大多数时候，知识只能通过对众多实验数据的慢慢积累得到。虽然这些实验都有各自的缺陷，但是他们总能聚合成为一个共识性的结论。

当今科学最令人激动的尝试和努力，就是追寻对人类行为本性的

理解。掌握这本书中的观念，将使你能够跟上追寻的脚步，或许还能真正成为其中一员。